TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Andreas Pung

Konvolutsioonilisel neurovõrgul põhineva teksti klassifitseerimismudeli interpreteerimine kliinilisel andmestikul

Bakalaureusetöö (9 EAP)

Juhendaja: Kairit Sirts, PhD

Tartu 2019

Konvolutsioonilisel neurovõrgul põhineva teksti klassifitseerimismudeli interpreteerimine kliinilisel andmestikul

Lühikokkuvõte:

Bakalaureusetöös interpreteeritakse konvolutsioonilist teksti klassifitseerimise neurovõrku: miks just niimoodi neurovõrk klassifikatsiooniotsuseid teeb. Analüüsi teostati kliinilisel DementiaBanki andmestikul, kus on Alzheimeri tõvega inimeste kirjeldused Bostoni küpsisevarguse fotost. Binaarse klassifikatsiooni ülesanne seisnes etteantud teksti põhjal klassi tuvastamises: kas isikul on Alzheimeri tõbi või mitte. Programmeeriti valmis Jacovi et al. artiklis (2018) kirjeldatud interpretatsioonimeetodid. Samuti interpreteeritakse konkreetseid tekste. Interpretatsioonimeetoditest andsid hea tulemuse informatiivsete ja ebainformatiivsete n-grammide leidmine, sõnede aktivatsioonivektorite leidmine ning nende klasterdamine. Vastandlike n-grammide analüüs andis halvema tulemuse andmestiku eripärade tõttu.

Võtmesõnad:

Neurovõrkude interpreteerimine, konvolutsioonilised neurovõrgud, masinõpe, teksti binaarne klassifitseerimine, DementiaBank, Alzheimeri tõbi

CERCS: P176 Tehisintellekt

Interpreting a Convolutional Text Classification Neural Network on a Clinical Dataset

Abstract:

In this Bachelor's Thesis, a convolutional text classification neural network is interpreted to find out why the neural network makes such predictions. To perform the analysis, the clinical DementiaBank dataset was used in which people with Alzheimer's disease describe the Boston cookie theft image. The task of the binary classification was to identify based on the given text whether a person has Alzheimer's or not. Interpretation methods described in Jacovi et al. (2018) were implemented. In addition to that, concrete examples of texts are interpreted in this thesis. Out of all the analyses performed, informative and uninformative ngrams and slot activation vectors with their clustering yield good results. Negative ngrams analysis results were substandard because of the specificity of the dataset.

Keywords:

Interpreting neural networks, convolutional neural networks, machine learning, binary text classification, DementiaBank, Alzheimer's disease

CERCS: P176 Artificial Intelligence

Sisukord

Sissejuhatus

Tänapäeva maailmas kasvab tehisintellekti roll oluliselt. Masinõpet kasutatakse väga erinevates tehnoloogilistes lahendustes ning selle vajalikkus ning olulisus on suurem kui kunagi varem. Neurovõrgud on väga olulised masinõppe meetodid, mis just viimasel ajal on erilist populaarsust kogunud. Näiteks teame, et mõne aasta eest otsustas Google Translate üle minna statistiliselt masintõlkelt neurovõrkudel põhinevatele masintõlke meetoditele (Turovsky, 2016).

Käesolevas bakalaureusetöös tegeletakse teksti binaarse klassifikatsiooni ülesandega. Klassifitseerimiseks kasutatakse masinõpet – neurovõrke. Täpsemalt kasutatakse konvolutsioonilisi neurovõrke. Püütakse aru saada, kuidas ning mis alustel konvolutsiooniline neurovõrk otsustab, millisesse klassi tekst klassifitseerida.

Palju on uuritud, kuidas konvolutsiooniline neurovõrk klassifitseerib pilte (Zeiler & Fergus, 2014). Teksti klassifitseerimise ülesannet sarnase arhitektuuriga neurovõrkude puhul on uuritud vähem. Konvolutsioonilisi neurovõrke teksti klassifitseerimisel on kasutatud näiteks filmiarvustuste meelsusanalüüsi teostades, küsimuse liiki klassifitseerides ja Twitteri säutsude meelsust tuvastades (Kalchbrenner, et al., 2014). Huvitav oleks teada saada, millised tulemused saadakse kliinilisel andmestikul. Käesolevas bakalaureusetöös kasutatakse DementiaBanki andmestikku, mis sisaldab Alzheimerit põdevate ning tervete inimeste (transkribeeritud) kirjeldusi Bostoni küpsisevarguse fotost. Iga tekst on märgendatud – kas tegemist on patsiendiga või kontrollgrupi isikuga.

Bakalaureusetöö põhiliseks allikaks on Alon Jacovi, Oren Sar Shalomi ning Yoav Goldbergi kirjutatud artikkel „Understanding Convolutional Neural Networks for Text Classification“ (2018). Artiklis on välja toodud meetodid konvolutsioonilise teksti klassifitseerimise neurovõrgu interpreteerimiseks. Täpsemalt, on välja pakutud analüüs informatiivsete ja ebainformatiivsete n-grammide tuvastamiseks, sõnede aktivatsioonivektorite leidmiseks ja nende klasterdamiseks ning vastandlike n-grammide leidmiseks. Samuti interpreteeritakse artiklis konkreetseid meelsusanalüüsi valdkonda kuuluvaid tekste – filmi- ja tootearvustusi. Konkreetseid tekste interpreteeritakse ka käesolevas bakalaureusetöös.

Bakalaureusetöö eesmärk on rakendada Jacovi et al. (2018) artiklis väljapakutud interpretatsioonimeetodeid teisel, kliinilisel andmestikul, et katsetada meetodite usaldusväärsust. Konvolutsioonilise neurovõrgu filtrite töö mõistmine aitab paremini analüüsida mudeli poolt tehtud vigu ning võib anda mõtteid mudeli edasiarenduseks. Eesmärgi teostamiseks

programmeeritakse artiklis välja toodud interpretatsioonimeetodid, rakendatakse mudelit ja meetodeid DementiaBanki andmestikul ning esitatakse süstemaatiliselt saadud tulemused.

Bakalaureusetöös püütakse vastata järgmisele uurimisküsimusele: kuivõrd rakendatavad on Jacovi et al. poolt välja pakutud meetodid kliinilisel andmestikul treenitud klassifitseerija interpreteerimisel?

Leiti, et informatiivsete ja ebainformatiivsete n-grammide, sõnede aktivatsioonivektorite leidmine koos klasterdamisega andsid hea tulemuse. Vastandlike n-grammide leidmine andis halvema tulemuse andmestiku eripärade tõttu.

Esimeses peatükis antakse lühiülevaade masinõppe olemusest ja neurovõrkudest. Teises peatükis kirjeldatakse artiklis (Jacovi, et al., 2018) välja toodud neurovõrgu interpreteerimise meetodeid. Kolmandas peatükis kirjeldatakse analüüsitavat DementiaBanki andmestikku lähemalt. Neljandas peatükis selgitatakse olulisi implementatsioonidetaile. Viies peatükk on analüüsiv – seal tuuakse välja töö uurimusliku osa tulemused.

Bakalaureusetöö autor soovib väga tänada oma juhendajat Kairit Sirtsu, kes andis autorile väga palju häid soovitusi ja nõuandeid.

1. Masinõppe ja neurovõrkude teoreetiline taust

Selles peatükis antakse lugejale vajalikud teoreetilised taustateadmised: milles seisneb teksti klassifitseerimine, masinõpe ning neurovõrgud. Lõpuks selgitatakse, millise arhitektuuriga neurovõrku käesolevas bakalaureusetöös rakendatakse.

1.1 Teksti klassifitseerimine

See alapeatükk põhineb Aggarwali ja Zhai raamatul (2012). Allika kohaselt on teksti klassifitseerimine protsess, kus tekstile seatakse vastavusse kindel märgend. Näiteks võidakse kategoriseerida uudiseid ning ka tekstide põhjal otsustada inimeste meelsuse üle – kas ollakse positiivselt või negatiivselt meelestatud mingi teema suhtes.

Üldiselt on olemas kaks erinevat teksti esitusviisi. Esimene on sõnahulk (ingl *bag-of-words*), kus teksti esitatakse hulgana koos infoga nende sageduse kohta. Selline esitusviis on peaaegu sõltumatu sõnade järjekorrast tekstis. Teine meetod on esitada teksti otse sõnedena, kus iga tekst on sõnede järjend.

Üks levinumaid teksti eeltöötlemise viise on stoppsõnade eemaldamine ning tüvestamine (sõna asendatakse sõnatüvega) või lemmatiseerimine (sõna asendatakse sõna algvormi ehk lemmaga). Stoppsõnade eemaldamisel leitakse teksti kõige levinumad sõnad, mis tüüpiliselt ei mõjuta klassifikatsiooniotsust.

Erinevaid teksti klassifitseerimise algoritme on mitmeid. Reeglipõhised klassifitseerijad klassifitseerivad tekste sissekodeeritud reeglite põhjal. Masinõppepõhiseid klassifitseerimisalgoritme on mitmeid. Nende hulka kuuluvad näiteks otsustuspuud, tugivektormasinad, Bayesi klassifitseerijad ning neurovõrkudel põhinevad klassifitseerijad, mis on kasutuses ka käesolevas bakalaureusetöös.

1.2 Masinõpe

Tänapäeval on tehisintellekt (ingl *artificial intelligence*) jõudsalt kasvav teadusharu, millel on palju praktilisi rakendusi ning uurimisteemasid. Püüeldakse intelligentse tarkvara poole, mis automatiseeriks rutiinseid ülesandeid, saaks aru kõnest või piltidest, teha meditsiinilisi diagnoose või toetaks üldiselt teadustegevust (Goodfellow, et al., 2016).

Sissekodeeritud teadmistel põhinevad tehisintellekti süsteemid ei toimi eriti hästi. See viitab sellele, et tehisintellekti süsteemidel on vaja võimet üldistada teadmisi toorandmetest ehk õppida – seda nimetataksegi masinõppeks (Goodfellow, et al., 2016).

Goldbergi õpiku (2017) kohaselt seisneb juhendatud masinõppe olemus uute mehhanismide loomises, mis näidete põhjal suudavad luua üldistusi. Näiteks on üheks traditsiooniliseks ülesandeks e-posti klassifitseerimine spämmiks ja mittespämmiks. Seejuures ei disainita konkreetset algoritmi, mis seda teeks, vaid luuakse üldine algoritm, mille sisendiks on hulk märgendatud näiteid spämmkirjadest ja tavalistest kirjadest. Algoritmi väljundiks on funktsioon või mudel, mis võtab sisendiks ühe konkreetse kirja ning väljundina tagastab märgendi: kas tegemist on spämmiga või mitte.

Formaalsemalt, öeldakse, et arvutiprogramm õpib kogemusest E mõnda ülesannete klassi T sooritusvõimega P, kui selle sooritus ülesandes T mõõdetuna sooritusvõime P põhjal paraneb läbi kogemuse E (Mitchell, 1997).

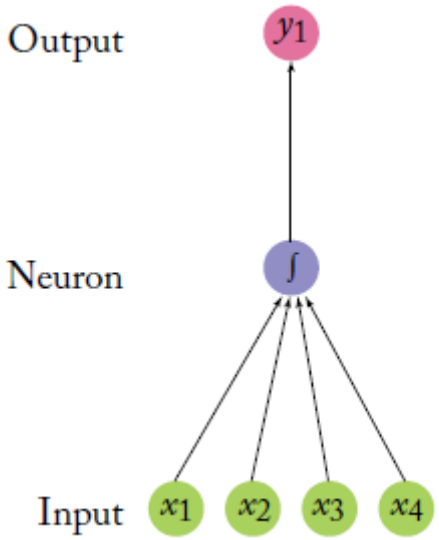
Klassifikatsioon on defineeritud Aggarwali ja Zhai õpiku (2012) kohaselt järgnevalt. Olgu meil treeningandmestik $\mathcal{D} = \{X_1, \dots, X_N\}$, nii et iga element on märgendatud kindla klassiga. Treeningandmestikule vastav klass on saadud elementide $\{1, \dots, k\}$ hulgast, kus erinevaid klasse on k tükki. Treeningandmestikku kasutatakse klassifikatsioonimudeli loomiseks, mis viib tekstile vastavusse konkreetse klassimärgendi. Treenitud mudel ennustab klassimärgendi etteantud testhulga tekstile, mille klass pole veel teada.

1.3 Neurovõrkude olemus

Goodfellow et al. õpiku (2016) kohaselt on sügavad pärilevivõrgud (ingl *feedforward network*), tuntud ka kui pärilevi tehisnärvivõrgud, neurovõrgud või mitmekihilised pertseptronid, ühed väga levinud masinõppe mudelid. Selgitatakse, et pärilevi neurovõrgu ülesandeks on ligilähedaselt hinnata mingit funktsiooni f^*. Klassifitseerija korral funktsioon $y = f^*(x)$ loob vastavuse sisendist x mingisse klassi y. Pärilevi neurovõrk defineerib kujutuse $y = f(x; \theta)$ ning õpib mudeli parameetrid θ, nii et funktsiooni väljund oleks võimalikult ligilähedane soovitud väljundile. Neid mudeleid kutsutakse pärilevi mudeliteks, sest info levib sisendist x edasi läbi vahepealsete arvutuste, mis defineerivad funktsiooni f, väljundisse y.

Goldbergi ja Hirsti õpiku (2017) kohaselt on neurovõrgud inspireeritud ajust, mis koosneb arvutuslikest üksustest, mida nimetatakse neuroniteks. Ühel neuronil on skalaarsed sisendid ning väljundid. Iga neuronil on oma kaal. Neuron korrutab iga sisendi selle kaaluga ning kõige tavalisemal juhul summeerib korrutised. Seejärel rakendatakse saadud tulemusele

mittelineaarset funktsiooni ning tulemus saadetakse väljundisse. Joonisel 1 on kujutatud ühte neuronit f, millel on neli sisendit (arvu) x_1, x_2, x_3, x_4 ning väljund y_1.



Joonis 1. Nelja sisendiga neuron (Goldberg, 2017).

Järgmises peatükis kirjeldatakse, kuidas neurovõrku treenitakse ning kuidas neurovõrk täpsemalt toimib.

1.4 Neurovõrkude treenimine

Järgnevalt refereeritakse Yann LeCun et al. artiklit „Deep learning“ (2015). Artikli kohaselt on neurovõrkude jaoks vaja suurt andmestikku näidisandmetega. Klassifikatsiooniülesande korral on treenimise käigus vaja, et neurovõrgu väljundina tagastatav korrektne klass oleks kõige suurema aktivatsiooniga ehk skooriga (väljundneuronil suurim sisendite ja kaalu korrutiste summa). Kasutusel on veafunktsioonid, mis iseloomustavad neurovõrgu väljundi ja tegeliku soovitava väljundi vahelist vea suurust. Et mudel toimiks paremini, treenitakse neurovõrku. Treenimise käigus muudetakse neuronite kaalusid, et minimeerida veafunktsiooni väärtust.

Kaalude korrektseks muutmiseks on vaja arvutada gradiendivektor. See näitab iga kaalu kohta, kui palju veafunktsiooni väärtus suureneb või väheneb, kui kaalu väga väikesel määral muuta. Kaalude vektorit muudetakse gradiendivektorile vastupidises suunas, sest veafunktsiooni väärtust soovitakse minimeerida. Praktikas kasutatakse enamjaolt optimiseerijana stohhastilist gradientlaskumist. Selle protsessi käigus arvutatakse keskmine gradient

väikese osa treeningnäidete kohta ning muudetakse vajalikul määral kaale. Protsessi korratakse paljude treeningandmete hulkadega kuni keskmine klassifikatsioonifunktsiooni väärtus enam ei muutu.

Peale treenimist hinnatakse mudeli headust täiesti eraldiseisval andmestikul – testandmestikul. Selle abil hinnatakse mudeli üldistusvõimet. Üheks mudeli headuse meetrikaks on täpsus – õigesti märgendatud klasside arvu ning terve testandmestiku suuruse jagatis.

Gradienti arvutatakse tagasilevi algoritmi abil. Selleks rakendatakse tuletiste ahelareeglit. Põhimõtteliselt töötab see niimoodi, et tuletist (ehk gradienti) on võimalik leida, hakates tagurpidi kihtide kaupa liikuma väljundist tagasi sisendini. Kui gradiendi funktsioonid on arvutatud, on lihtne leida konkreetse kaalu gradient.

Nagu eelnevalt mainitud, siis ühest kihist teise liikumisel summeeritakse neuronis eelmise kihi neuronitest saadud väärtused, mis on korrutatud praeguse neuroni kaaluga. Seejärel antakse tulemus mittelineaarse funktsiooni argumendiks. Hetkel on üheks populaarseimaks selliseks funktsioon mittenegatiivne lineaarfunktsioon $ReLU(x) = max(x, 0)$.

Et neurovõrk treeningandmestikku pähe ei õpiks, kasutatakse väljajätumeetodit (ingl *dropout method*). Väljajätumeetod seisneb treenimise ajal teatud neuronite ja nende ühenduste nullimises, parameetriks on võimalik anda, mitu protsenti terve neurovõrgu neuroneid nullitakse (Srivastava, et al., 2014).

Treenimise käigus kasutatakse tavaliselt rohkem kui ühte ja vähem kui kõiki treeningnäiteid korraga – seda nimetatakse miniplokktreeninguks (ingl *mini-batch training*) (Goodfellow, et al., 2016). Lisaks väljajätumeetodile võib mõjutada neurovõrgu treenimist ka ploki suurus (ingl *batch size*), mis näitabki seda, mitu treeningnäidet korraga vaatluse all on.

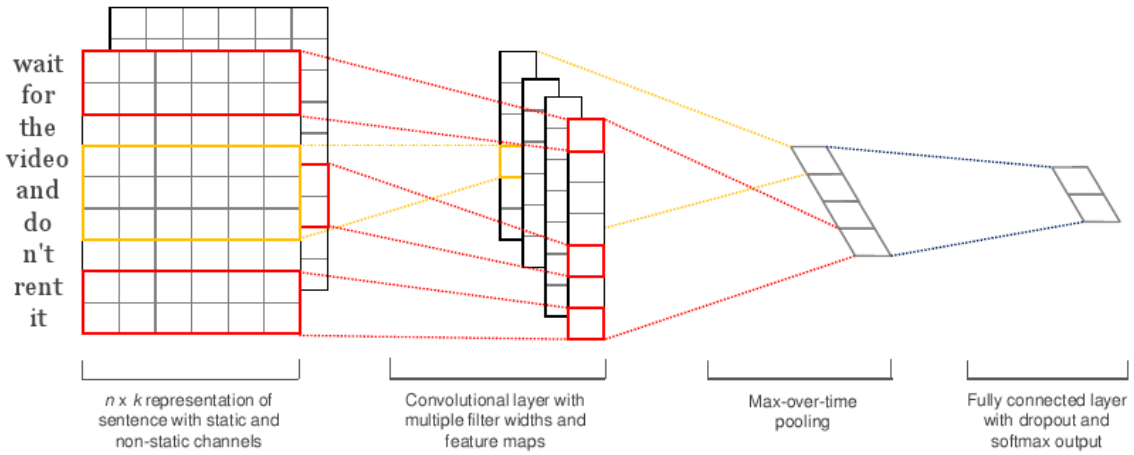
1.5 Konvolutsioonilise teksti klassifitseerimise neurovõrgu üldine arhitektuur

LeCuni (2015) põhjal koosnevad konvolutsioonilised neurovõrgud konvolutsioonilistest ning ahenduskihtidest (ingl *pooling layer*). Konvolutsioonilises kihis on ühikud struktureeritud tunnuskaartidena (ingl *feature map*). Iga kiht koosneb filtritest, igaühel neist on oma kindlad kaalud. Ahenduskihi mõte on semantiliselt lähedasi tunnuseid ühendada.

Ühe populaarseima konvolutsioonilise teksti klassifitseerimise neurovõrgu arhitektuuri pakkus välja Yoon Kim oma artiklis „Convolutional Neural Networks for Sentence Classification“ (2014). Väljapakutud arhitektuur on esitatud joonisel 2. Arhitektuuri lühikirjeldus on refereeritud samast artiklist.

Sõnavektorite (ingl *embedding vector*) algväärtustamine neurovõrgust saadud vektoritega (eeltreenitud vektoritega), on populaarne meetod, mis aitab parandada mudeli täpsust, kui pole suurt treeningandmestikku (Collobert, et al., 2011). Sõnavektoreid kasutab ka Kim.

Terve lause on esitatud maatriksina, kus eeltreenitud sõnavektorid on omavahel konkateneeritud (joonisel vasakul, $n \times k$ maatriks). Seejärel rakendatakse maatriksile konvolutsiooni operatsiooni. Kindlaks on määratud akna suurus – mitut sõne korraga vaadeldakse. Aken „liugleb“ üle kõigi sõnede ning ühe filtri kaalud korrutatakse vaatluse all oleva maatriksi „liugakna“ osaga. Samuti liidetakse juurde vabaliige ja rakendatakse mittelineaarset aktivatsioonifunktsiooni. Kui kõikide konvolutsiooniliste kihtide kõik filtrid on oma tulemuse vektorisse talletanud, siis rakendatakse *max-poolingu* operatsiooni, et vastava filtri vektorist saada kõige suurema aktivatsiooniga n-gramm (n-sõnest koosnev sõneennik). Kõige suurema aktivatsiooniga n-gramm peaks olema klassifikatsiooniotsuse poolest kõige suurema tähtsusega. Kõige viimasena saadetakse *max-poolitud* vektor ühte täissidusasse neurovõrgu kihti, mille väljund omakorda liigub *softmax*-kihti (normaliseeritud eksponentfunktsioon), et väljastataks klasside tõenäosusjaotus.



Joonis 2. Yoon Kimi (2014) väljapakutud konvolutsioonilise teksti klassifitseerimise neurovõrgu arhitektuur.

Järgmises alapeatükis kirjeldatakse konkreetselt siin töös kasutatavat arhitektuuri.

1.6 Töös kasutatava neurovõrgu arhitektuuri kirjeldus

See alapeatükk põhineb Jacovi et al. (2018) artiklis välja toodud arhitektuuril. Bakalaureusetöö eesmärk on rakendada artiklis väljapakutud interpretatsioonimeetodeid teisel, kliinilisel andmestikul, et katsetada meetodite usaldusväärsust.

Artiklis käsitletakse üldlevinud arhitektuuri, kus iga teksti sõne on esitatud sõnavektorina, sellele rakendatakse konvolutsioonilist kihti m filtriga, mille väljundiks on iga teksti n-grammi kohta ka m-mõõtmeline vektor. Vektorid kombineeritakse, kasutades *max-poolingut,* millele järgneb kohe mittenegatiivne lineaarfunktsioon (ReLu). Saadud vektor suunatakse edasi lineaarsesse kihti lõpliku klassifikatsiooni jaoks.

Olgu meil n sõna pikkune sisendtekst $w_1, \dots, w_n$. Iga sõne esitatakse sõnavektorina, millel on d dimensiooni. Tegevuse tulemusena saadakse sõnavektorid $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n \in \mathbb{R}^d$. Saadud $d \times n$ maatriks läheb edasi konvolutsioonilisse kihti, kus „liugaken" asetatakse tekstile. Iga n-grammi pikkusega l kohta saame, et $\boldsymbol{u_i} = [\boldsymbol{w}_i, \dots, \boldsymbol{w}_{i+l-1}] \in \mathbb{R}^{d \times l}; 0 \leq i \leq n - l$.

Iga filtri $\boldsymbol{f_j} \in \mathbb{R}^{d \times l}$ kohta arvutatakse skalaarkorrutis $\langle \boldsymbol{u}_i, \boldsymbol{f}_j \rangle$. Konvolutsiooni rakendamise tulemusena saadakse maatriks $\boldsymbol{F} \in \mathbb{R}^{n \times m}$. Seejärel rakendatakse *max-poolingu* operatsiooni, misjärel saadakse vektor $\boldsymbol{p} \in \mathbb{R}^m$. Järgnevalt rakendatakse vektori $\boldsymbol{p}$ elementidele mittenegatiivset lineaarfunktsiooni. Viimase etapina väljastab lineaarne täissidus kiht $\boldsymbol{W} \in \mathbb{R}^{c \times m}$ klassifikatsiooniklasside tõenäosusjaotuse ning kõige tõenäolisem klass väljastatakse neurovõrgu poolt. Kokkuvõtvalt:

$$\boldsymbol{u_i} = [\boldsymbol{w}_i, \dots, \boldsymbol{w}_{i+l-1}]$$

$$F_{ij} = \langle \boldsymbol{u}_i, \boldsymbol{f}_j \rangle$$

$$p_j = ReLU(\max F_{ij})$$

$$\boldsymbol{o} = softmax(\boldsymbol{W}\boldsymbol{p})$$

Käesolevas bakalaureusetöös kasutatakse uni-, bi-, tri- ja neligramme ehk $l \in \{1, 2, 3, 4\}$. Nelja konvolutsioonilist kihti kasutatakse üksteisega paralleelselt ning konkateneeritakse vektorid $\boldsymbol{p}$, mis on saadud igat liiki n-grammi kohta.

2. Neurovõrgu interpreteerimise meetodid

Selles peatükis selgitatakse, millised neurovõrkude interpreteerimise meetodeid on kasutatud ning seejärel kirjeldatakse konkreetselt käesolevas bakalaureusetöös rakendatavaid meetodeid: informatiivsed ja ebainformatiivsed n-grammid, sõnede aktivatsioonivektorid ja nende klasterdamine ning vastandlikud n-grammid.

2.1 Võimalikud interpreteerimismeetodid

Üheks oluliseks interpretatsioonimeetodiks on hierarhiliste tähelepanuvõrkude kasutamine, kus teksti lauseid esitatakse hierarhilise struktuurina ning arvestatakse sõnade ja lausete konteksti, sest need võivad olla erineva informatiivsusega (Yang, et al., 2016). Sarnase arhitektuuriga on tehtud ka muid töid. Näiteks teksti klassifitseerimiseks on kasutatud arhitektuuri nimega *self-attention network*, kus on kajastatud kõikide teksti sõnade omavaheliste paaride seoseid (Letarte, et al., 2018). Selliste arhitektuuridega neurovõrkude interpreteerimist saab teha väga hästi, visualiseerides tähelepanu: näiteks seatakse tekstile vastava läbipaistvusega taustavärv, mis sõltub sellest, kui palju sõnad klassifikatsiooniotsust mõjutasid (Yang, et al., 2016; Letarte, et al., 2018).

On väidetud ka, et populaarsed neurovõrkude interpreteerimise meetodid ei paku ka lihtsale lineaarsele mudelile õiget selgitust (Kindermans, et al., 2018). Artiklis näidatakse, et lineaarses mudelis on kaalude olemuslik eesmärk vähendada müra sisendandmetes ning seega kaalud näitavad, kuidas andmetest signaali välja lugeda. See aga ei iseloomusta signaali ennast – seega peaksid interpretatsioonimeetodid liikuma kaalude vaatlemisest edasi. Autorid pakuvad välja PatternNeti ning PatternAttributioni meetodid, mis õpivad neurovõrgu selgitusi andmetest. Täpselt neidsamu eelmainitud meetodeid on rakendatud ka konvolutsioonilisele teksti klassifitseerimise neurovõrgule (Harbecke, et al., 2018).

On ilmunud ka artikkel, kus interpreteeriti DementiaBanki andmestikul treenitud konvolutsioonilist ning konvolutsioonilise võrgu kombinatsioone rekurrentse ja pika lühiajalise mälu neurovõrguga (Karlekar, et al., 2018). Artiklis interpreteeritakse neurovõrku aktivatsioonide klasterdamise meetodiga. Meetod seisneb neuronite aktivatsioonide käsitlemises ruumi koordinaatidena. Seejärel rakendatakse k-keskmiste klasteranalüüsi. Selle põhjal tekkisid omavahel sarnaste aktivatsioonidega klastrid. Artiklis kasutati veel ka teistsugust interpreteerimismeetodit – *first derivative saliency heat map*. See meetod aitab visualiseerida, millised sisendteksti sõnad mõjutavad klassifikatsiooni kõige rohkem.

Järgnevalt vaatleme konkreetseid interpretatsioonimeetodeid, mida käesoleva bakalaureusetöö uurimuslikus osas rakendatakse.

2.2 Informatiivsed ja ebainformatiivsed n-grammid

Käesolev ning järgmised alapeatükid põhinevad Jacovi et al. (2018) artiklis kirjeldatud interpretatsioonimeetoditele.

Olgu ahenduskihi läbinud n-grammid, mis satuvad vektorisse $\boldsymbol{p}$ tähistatud hulgana $S_{\boldsymbol{p}}$. N-grammid, mis ei ole hulgas $S_{\boldsymbol{p}}$, ei mõjuta järelikult klassifikatsiooni tulemust.

Hulga $S_{\boldsymbol{p}}$ n-grammid jaotuvad kahte klassi: tahtlikud ning juhuslikud. Tahtlikud n-grammid satuvad hulka $S_{\boldsymbol{p}}$, sest filter seadis n-grammile vastavusse kõrge aktivatsiooni tõenäoliselt selle tõttu, et nad on lõpliku klassifikatsiooniotsustuse tõttu informatiivsed. Juhuslikud n-grammid satuvad hulka $S_{\boldsymbol{p}}$ vaatamata sellele, et neil on madal aktivatsioon, sest mitte ükski teine n-gramm ei saavutanud kõrgemat aktivatsiooni. Need n-grammid tõenäoliselt lõplikku klassifikatsiooniotsust ei mõjuta. Selle analüüsi tulemusena püütakse eraldada tahtlikud n-grammid juhuslikest.

Oletatakse, et igal filtril on mingi kindel lävi, kus lävest suuremad väärtused viitavad informatiivsetele n-grammidele ning lävest madalamad väärtused ebainformatiivsetele, mis võib klassifitseerimise seisukohast kõrvale jätta. Järelikult otsitakse läve, mis eristaks kahte klassi. Klass, millele filter $\boldsymbol{f_j}$ panustab, on $c_j = argmax_k W_{kj}$. Nimetame klassi c_j filtri f_j klassiidentiteediks.

Klassifitseerija töö tulemusena saadakse vektorid $\boldsymbol{p^i}$ ning kogu neurovõrgu ennustused c^i. Iga filtri j kohta vaadeldakse väärtust $\boldsymbol{p}_j^i$ ning kas $c^i = c_j$. Iga filtri kohta saame andmestiku $\left(p_j^1, c^1 = c_j\right), \dots, \left(p_j^D, c^D = c_j\right)$. Püütakse leida läve t_j, mis eristaks väärtusi p_j^i juhtumitest, kus $c^i = c_j$ nendest, kus $c^i \neq c_j$. Saadakse järgmine lävede ja tõeväärtuste paar:

$$(X, Y)_j = \{\left(p_j^i, c^i = c_j\right) \mid j < m, i < D\}$$

Kui $p_j^i > t_j$, siis klassifitseerija ennustus ühtib filtri märgendiga, vastasel juhul ennustus ei ühti. Kuna praktikas ei ole eelnimetatud hulk lineaarselt eralduv, siis võetakse kasutusele uus mõiste – filtri ja läve kombinatsiooni puhtus. Puhtus on defineeritud kui informatiivsete n-grammide osakaal, mille aktivatsioon oli filtri lävest kõrgem. Olgu meil lävede andmestik (X, Y). Siis saame leida puhtuse niimoodi:

$$puhtus(f,t) = \frac{\left|\{(x,y) \in (X,Y)_f | x \geq t, y\ on\ tõene\}\right|}{\left|\{(x,y) \in (X,Y)_f | x \geq t\}\right|}$$

Heuristiliselt tuleks määrata filtri lävi nii madalaks, et saaks niivõrd kõrge puhtuse kui võimalik.

Välja pakutud lävede ja selle toimimismehhanismi usalduses veendumiseks tuleks n-grammid, mille aktivatsioon ei ületa läve, kõrvale jätta. Praktikas on see sama operatsioon, mis mittenegatiivse lineaarfunktsiooni asendamine uue lävefunktsiooniga:

$$lävi(x,t) = \begin{cases} x, kui\ x \geq t \\ 0\ vastasel\ juhul \end{cases}$$

Juhul, kui väljapakutud meetod läve kasutamiseks peab paika, ei tohiks klassifitseerimise täpsus testhulgal väheneda.

2.3 Sõnede aktivatsioonivektorid

Originaalartiklis (Jacovi, et al., 2018) nimetatakse meetodit „pilude“ aktivatsioonivektoriteks (ingl *slot activation vectors*), kuid selline toortõlge ei ole eesti keele pärane ning käesolevas bakalaureusetöös nimetatakse mõistet sõnede aktivatsioonivektoriteks.

Iga n-grammi $\boldsymbol{u_i} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_l]$ ja iga filtri $\boldsymbol{f}$ kohta arvutatakse skalaarkorrutis $\langle \boldsymbol{u}, \boldsymbol{f} \rangle$ (vaata peatükist 1.6). N-grammi aktivatsioon omakorda koosneb sõnede aktivatsioonide summadest. Individuaalne n-grammi aktivatsioon on ühe sõne sõnevektori $\boldsymbol{w_i}$ ja filtri kaalude osalise maatriksi $\boldsymbol{f}$ skalaarkorrutis:

$$\langle \boldsymbol{u}, \boldsymbol{f} \rangle = \sum_{i=0}^{l-1} \langle \boldsymbol{w_i}, \boldsymbol{f_{id:i(d+1)}} \rangle$$

Osaline maatriks $\boldsymbol{f_{id:i(d+1)}}$ on filtri kaalude veeruvektor, mis on positsioonil i ning mida tähistatakse $\boldsymbol{f(i)}$.Selle asemel, et leida skalaarkorrutiste summa ja leida n-grammi koguaktivatsioon, saab neid liidetavaid interpreteerida otse – öeldakse, et $\langle \boldsymbol{w_i}, \boldsymbol{f(i)} \rangle$ näitab, kui palju positsioon *i* filtri kaaludes $\boldsymbol{f}$ on aktiveeritud n-grammi positsioonil i oleva sõne poolt.

Nüüd saab teostada põhjalikumat analüüsi: n-grammi ja filtri paaride $\langle \boldsymbol{u} := [\boldsymbol{w}_1; \dots; \boldsymbol{w}_l], \boldsymbol{f} \rangle$ asemel vaadeldakse sõnede aktivatsioonivektorit $(\langle \boldsymbol{w_1}, \boldsymbol{f(1)} \rangle, \dots, \langle \boldsymbol{w}_l, \boldsymbol{f}(l) \rangle)$. Sõnede aktivatsioonivektor väljendab, kui palju iga n-grammi üksik sõne panustab terve n-grammi aktivatsioonile.

2.4 Sõnede aktivatsioonivektorite klasterdamine

Artiklis (Jacovi, et al., 2018) on välja pakutud hüpotees, et iga filter tuvastab mitme erineva semantilise klassiga n-gramme ning igal klassil on mõned domineerivad (märgatavalt kõrgema aktivatsiooniga) ning mittedomineerivad sõned.

Klasterdamisel klasterdatakse filtri läve ületavaid n-grammide aktivatsioonivektoreid (vaata peatükist 2.3). Kasutatakse algoritmi *mean shift clustering* (Fukunaga & Hostetler, 1975; Cheng, 1995). Algoritmis ei ole eelnevalt vaja määrata klastrite arvu ning algoritm ei tee eelduseid klastrite kuju kohta. Iga klaster tuvastab erineva sõnede aktivatsioonimustri. Klastri tsentroidiks on kõige tüüpilisem sõnede aktivatsioonivektor.

2.5 Vastandlikud n-grammid

Artiklis (Jacovi, et al., 2018) on välja pakutud ka teine hüpotees: sõne aktivatsioon ei pruugi olla maksimeeritud mitte selle pärast, et tuvastada selle olemasolu, vaid just mitteeksisteerimist, nii et teatud sõnad ei leiduks n-grammides. Neid nimetatakse vastandlikeks n-grammideks (ingl *negative ngrams*).

Vastandlike n-grammide leidmiseks otsitakse n-gramme, mis on esialgsete n-grammide „ümberpööratud“ versioonid. Olgu antud n-gramm $\boldsymbol{u}$, millele filter $\boldsymbol{f}$ seadis vastavusse kõrge aktivatsiooni. Otsime madala aktivatsiooniga n-gramme $\boldsymbol{u}'$, nii et n-grammide $\boldsymbol{u}$ ning $\boldsymbol{u}'$ vaheline Hammingi kaugus oleks võimalikult väike. Hammingi kauguseks nimetatakse n-grammi sõnede arvu, mida peab asendama, et saada ühest n-grammist teine (Hamming, 1950). Näiteks on trigrammide „ma vihkan seda“ ning „ma armastan seda“ Hammingi kaugus üks.

3. DementiaBanki andmestik

Selles peatükis kirjeldatakse lühidalt töös kasutatavat andmestikku.

Andmestikuna kasutati DementiaBankist saadud andmeid (Becker, et al., 1994). DementiaBank[1] on osa Talkbanki[2] korpusest. Andmestik sisaldab dementsete (Alzheimerit põdevate) ja tervete inimeste (kontrollgrupi isikute) kirjeldusi Bostoni küpsisevarguse (Goodglass & Kaplan, 1983) fotost (vaata joonist 3).

Joonis 3. Bostoni küpsisevarguse foto (Goodglass & Kaplan, 1983).

Andmestik koosneb CHAT formaadist (MacWhinney, 2000) transkribeeritud foto kirjeldustest. Andmed on jaotatud treening-, valideerimis- ning testandmestikuks. Andmed on esitatud sõnestatud kujul, millele järgneb märgend – kas tegemist on patsiendiga (P) või kontrollgrupi isikuga (C).

[1] https://dementia.talkbank.org/

[2] https://talkbank.org/

Tabel 1. DementiaBanki andmestiku olulisemad statistilised näitajad.

	Treeningandmestik	Valideerimisandmestik	Testandmestik
Tekstide kogus	296	101	101
Sõnastiku suurus	1321	818	767
Keskmine teksti pikkus	123.77	133.20	122.85
Teksti pikkuste standardhälve	63.10	63.66	76.81

Tabelis 1 on välja toodud DementiaBanki treening-, valideerimis- ja testandmestike olulisemad statistilised näitajad. Tähele tasub panna, et andmestik on väga väike. Tekste on kokku 498 tükki. Interpreteerimise analüüsis kasutati treening- ja valideerimisandmestiku konkateneeritud andmestikku, sest nende andmestike põhjal oli mudel treenitud või sai treenimise käigus tagasisidet.

4. Mudeli implementatsioon

Selles peatükis selgitatakse lühidalt, kuidas saadi neurovõrgu mudel ning milliseid hüperparameetreid mudelis kasutati.

4.1 Implementatsioonidetailid

Konvolutsiooniline teksti klassifitseerimise neurovõrk programmeeriti keeles Python 3. Selleks kasutati teeke PyTorch[3] ning TorchText[4]. Mudel programmeeriti valmis Ben Trevetti PyTorchi meelsusanalüüsi õpetuse põhjal, mis on kättesaadav GitHubis[5]. Interpretatsioonimeetodite baaskoodi sai uurimistöö autor juhendajalt.

Hüperparameetrite valimise põhjalikum kirjeldus on esitatud alapeatükis 4.2. Mudel implementeeriti järgmiste hüperparameetritega:

1) eeltreenitud sõnavektorid glove.6B.100D;
2) neli konvolutsioonilist kihti – uni-, bi-, tri- ning neligrammide kihid;
3) väljajätumeetod 0,4 – 40% neuronite aktivatsioonidest seatakse nulliks;
4) ploki suurus 32 teksti;
5) Adami optimiseerija;
6) 10 filtrit iga kihi kohta (seega kokku tervel võrgul 40 filtrit);
7) binaarne ristentroopia kahjufunktsioon koos sigmoidiga.

Järgnevalt selgitatakse detailsemalt, kuidas just sellised hüperparameetrid välja valiti.

4.2 Hüperparameetrite valimine

Hüperparameetrite väljavalimiseks optimeeriti sõnavektoreid, ploki suurust, vaadeldi, kas analüüsi hulka lisada unigramme või mitte ning väljajätumeetodit (lühendatud DO). Selleks prooviti läbi mõningaid hüperparameetrite kombinatsioone. Mudeli headust hinnati valideerimishulgal.

[3] https://pytorch.org/

[4] https://github.com/pytorch/text

[5] https://github.com/bentrevett/pytorch-sentiment-analysis

Tabel 2. Hüperparameetrite valimise protsess.

Sõnavektorid	Ploki suurus	Filtrite suurused	DO	Treen-kadu	Treen-täpsus	Val-kadu	Val-täpsus	Test-kadu	Test-täpsus
glove.6B.100d	64	2–4	0.5	0.08	98.75	0.44	85.43	0.62	72.78
glove.6B.100d	64	1–4	0.5	0.03	99.50	0.48	86.21	0.64	76.27
glove.6b.50d	64	1–4	0.5	0.39	84.94	0.52	78.61	0.58	67.38
glove.6b.200d	64	1–4	0.5	0.23	91.25	0.44	84.44	0.54	72.42
glove.6b.100d	296	1–4	0.5	0.36	87.50	0.52	80.20	0.56	71.29
glove.6b.100d	32	1–4	0.5	0.14	96.56	0.34	85.94	0.52	80.16
glove.6b.100d	16	1–4	0.5	0.05	99.34	0.42	85.71	0.66	66.43
charngram.100d	32	1–4	0.5	0.34	87.50	0.48	82.81	0.56	66.56
glove.6b.100d	32	1–4	0.3	0.01	100.00	0.42	85.16	0.60	78.91
glove.6b.100d	32	1–4	0.7	0.23	92.50	0.40	85.94	0.54	77.03
glove.6b.100d	**32**	**1–4**	**0.4**	**0.02**	**99.38**	**0.40**	**87.50**	**0.56**	**82.03**

Tabelis 2 on välja toodud hüperparameetrite valimise protsess. Esmalt vaadeldi, kas unigrammide lisamine eraldi konvolutsioonilise kihina annab positiivse efekti või mitte. Unigrammide lisamine andis valideerimistäpsuseks 86,21%, mis on parem tulemus kui ilma unigrammideta (85,43%), seega jätkati koos unigrammidega.

Järgnevalt vaadeldi klassifikatsioonitäpsuseid sõnavektoritega Glove (Pennington, et al., 2014), täpsemalt vaadeldi vektorite dimensionaalsusi. Parima tulemuse andsid 100-dimensioonilised vektorid, 86,21%. 50-dimensiooniliste vektorite kasutamisel saadi täpsuseks 78,61% ning 200-dimensiooniliste korral 84,44%. Järgnevalt püüti muuta ploki suurust. Prooviti nii täisandmestiktreeningut (kus ploki suurus on võrdne treeningtekstide arvuga), mis andis valideerimistäpsuseks 80,20%; ploki suurust 32, mis andis tulemuseks 85,94% kui ka ploki suurust 16, millega saavutati valideerimistäpsus 85,71%. Siiski otsustati edasi minna ploki suurusega 32 (85,94%) ja mitte ploki suurusega 64 (parim valideerimistäpsus seni 86,21%) pragmaatilistel põhjustel, sest treenimine toimus palju kiiremini. Prooviti ka täiesti teisi sõnavektoreid, nimelt 100 dimensiooniga charNgrami (Hashimoto, et al., 2017), millega saadi valideerimistäpsus 82,81% – seega kasutati edasi sõnavektoreid Glove. Kõige viimasena häälestati väljajätumeetodit. Väärtus 1 näitab, et kõikide neuronite aktivatsioonid nullitakse, 0 seevastu, et ühtegi aktivatsiooni ei nullita. Prooviti väärtuseid 0,3, millega saadi valideerimistäpsus 85,16%; 0,7, millega saadi valideerimistäpsus 85,94% ning lõpuks väärtust 0,4, millega saadi valideerimistäpsuseks 87,50%. Kõige paremaks testimistäpsuseks varjatud testhulgal saadi 82,03%.

Tabel 3. Väljavalitud parameetritega mudeli täpsuste ja kadude statistikud.

N = 100 treenimist	Keskmine	Standardhälve	Miinimum	Maksimum
Treenimistäpsus	93.98%	7.36%	59.69%	99.69%
Treenimiskadu	0.189	0.144	0.021	0.672
Valideerimistäpsus	86.24%	1.86%	82.03%	91.41%
Valideerimiskadu	0.391	0.062	0.297	0.669
Testimistäpsus	73.25%	4.40%	63.91%	82.81%
Testimiskadu	0.555	0.054	0.413	0.679

Täpsuste usalduses veendumiseks treeniti mudelit 100 korda ning leiti treening-, valideerimis- ja testhulkade aritmeetilised keskmised, standardhälbed, miinimumid ja maksimumid. Tabelis 3 on välja toodud saadud tulemused. Keskmiseks testimistäpsuseks saadi 73,25% standardhälbega 4,40%.

Klassifitseerimise täpsus on võrreldav ka teiste tulemustega. Uurimuses, kus kasutati juhuslike metsade klassifitseerijat ja kauguspõhiseid, ideetiheduse ja -tõhususe ning leksikosüntaktilisi ning akustilisi tunnuseid, saavutati täpsuseks 80% (Yancheva & Rudzicz, 2016).

Ideetiheduse meetodit kasutades saadi logistilise regressiooniga ühes uurimuses täpsuseks 72,1% standardhälbega 0,6% (Sirts, et al., 2017).

5. DementiaBanki andmestikul treenitud neurovõrgu interpreteerimine

Selles peatükis esitatakse saadud tulemused erinevate analüüside kaupa: informatiivsed ja ebainformatiivsed n-grammid, sõnede aktivatsioonivektorid ja nende klasterdamine, vastandlikud n-grammid ning kõige viimasena interpreteeritakse konkreetseid tekste.

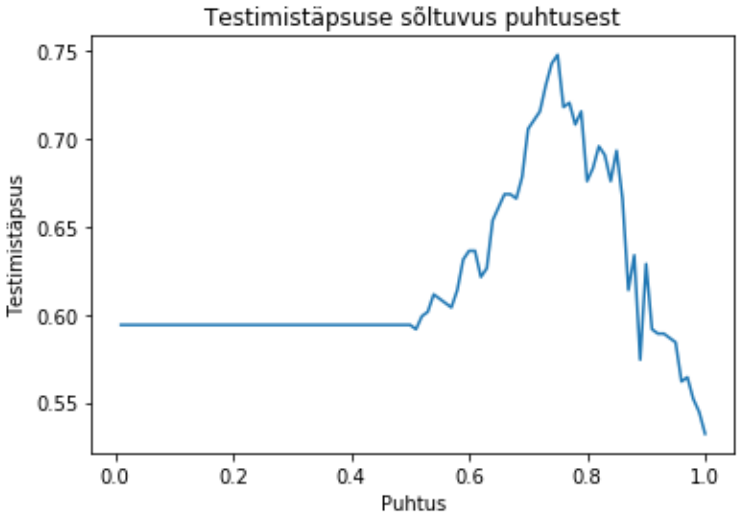
Analüüside teostamiseks kasutatud Pythoni programmi lähtekood on avalikult kättesaadav GitHubis[6].

5.1 Informatiivsed ja ebainformatiivsed n-grammid

Esimesena teostati analüüsi filtrite kohta: millised peavad olema filtrite läved, et eristada informatiivseid n-gramme ebainformatiivsetest.

Esmalt analüüsitakse, milline universaalne puhtuse väärtus on andmestikul sobivaim. Selleks püüti leida, millise puhtuse väärtuse juures saavutatakse kõige parem testimistäpsus. Selleks arvutati iga puhtuse väärtuse $(0, 0.01, \ldots, 0.99, 1.00)$ kohta iga filtri lävi. Seejärel asendati neurovõrgu originaalarhitektuuris mittenegatiivne lineaarfunktsioon lävefunktsiooniga. Funktsioon selekteerib välja ainult läve ületavad n-grammid. Selle põhjal arvutati testimistäpsused erinevate puhtuste väärtuste juures. Joonisel 4 on kujutatud DementiaBanki andmestiku testimistäpsuse sõltuvus puhtusest. Leiti, et optimaalne puhtus on 0,75 – selle puhtuse väärtuse juures saavutati testimistäpsuseks 0,75. Võrreldes Jacovi et al. saadud tulemustega, muutus testimistäpsus palju rohkem, vahemikus 0,59 kuni 0,75. Võib järeldada, et teatud (ebainformatiivsete) n-grammide ärajätmine lävefunktsiooni abiga võib väga suurel määral mõjutada testhulga täpsust – järelikult ka teatud tekstide klassifikatsiooniotsust. Seega, kuna testimistäpsuse ulatus oli niivõrd suur, võib oletada, et korrektse läve leidmine on väga oluline.

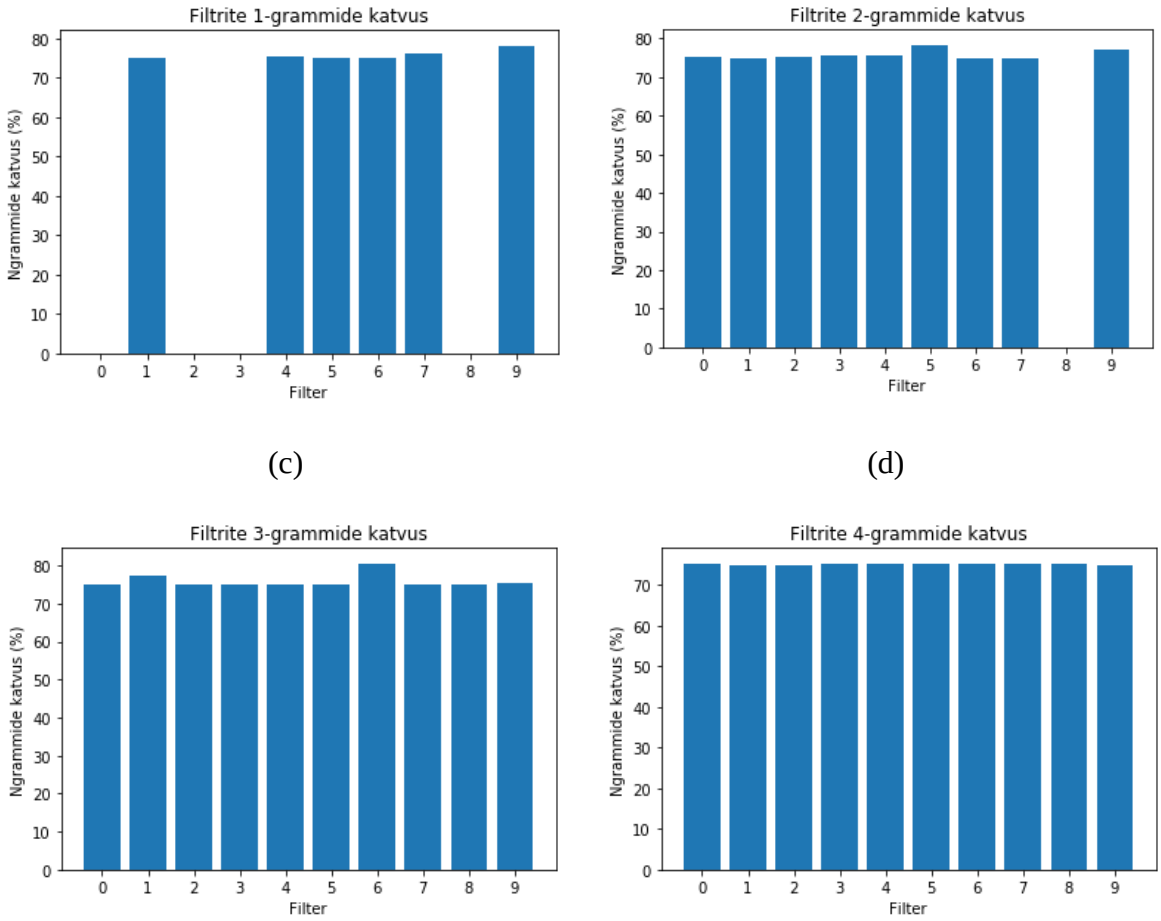
[6] https://github.com/andreaspung/interpreting-db



Joonis 4. Testimistäpsuse sõltuvus puhtusest.

Järgnevalt vaadeldakse puhtuse väärtuse 0,75 juures n-grammide katvust iga filtri kaupa. Katvus on läve ületanud (informatiivsete) filtri klassi ja neurovõrgu otsustusega ühtivate n-grammide arvu ning kõikide läve ületanud n-grammide jagatis. Joonisel 5 on kujutatud uni-, bi-, tri- ja neligrammide kõiki kümmet filtrit ning katvust protsentides. Leiti, et unigrammidel pidas mudel informatiivseks ainult kuut filtrit. See tähendab, et nelja filtri väljavalitud *max-poolitud* n-grammid jäeti vaatluse alt välja, sest need ei ületanud läve. See võib olla selgitatav sellega, et mingi ühe üksiku sõne olemasolu ei pruugi väga kindlalt eristada kontrollgrupi indiviidi patsiendist. Bigrammide korral peeti ainult ühe filtri väärtuseid ebainformatiivseks, selle filtri läve (ilma vabaliiget juurde liitmata) ületas ainult üks bigramm: „the stool“. Kui vaadelda neid n-gramme, mis läve kindlalt ei ületa, siis leiame näiteks bigrammi „a stool“. Intuitiivselt mõeldes see aga ei tundu eriti loogiline olevat, sest artiklite „a“ ja „the“ vahel on väga väike tähenduslik erinevus ning see ei tohiks mõjutada klassifikatsiooniotsust. Samuti ei ületanud filtri läve mitmed bigrammid, mis sisaldasid keerukamaid sõnu nagu näiteks „climbing“, „interpreted“, „daydreaming“ ning „depression“. Järelikult peaks suhtuma selle konkreetse filtri lävesse väga kriitiliselt. Tri- ja neligrammide korral olid lõpliku klassifikatsiooniotsustuse juures kasutusel kõikide filtrite n-grammid. Nende filtrite korral, mida mudel pidas vastava läve juures informatiivseks, ületasid läve 70 – 80 % kõikidest n-grammidest.

(a) (b)



(c) (d)

Joonis 5. Filtrite uni- (a) kuni neligrammide (d) katvus iga filtri kaupa.

Järgnevalt vaadeldakse ülevaadet igast filtrist: mis klassi filtrid tuvastavad, mis on nende lävi ning kui palju n-gramme läve ületavad.

Tabel 4. Ülevaade igast filtrist.

Unigrammid					**Trigrammid**				
Filter	Klass	Lävi	Õigeid	Kokku	Filter	Klass	Lävi	Õigeid	Kokku
0	P	3.038	0	0	0	P	1.278	121	163
1	C	0.720	141	190	1	C	0.935	157	212
2	P	3.177	0	0	2	P	1.106	147	198
3	P	2.236	0	0	3	C	0.289	187	252
4	P	0.957	142	188	4	C	1.675	100	134
5	C	0.223	177	239	5	C	0.534	163	220
6	P	0.571	163	219	6	P	1.446	141	190
7	C	0.702	150	201	7	C	1.310	136	183
8	C	1.074	0	0	8	P	1.817	97	129
9	P	2.316	43	55	9	C	0.098	165	222

Bigrammid					Neligrammid				
Filter	Klass	Lävi	Õigeid	Kokku	Filter	Klass	Lävi	Õigeid	Kokku
0	P	0.896	139	187	0	P	0.314	168	227
1	P	0.895	114	153	1	C	0.300	185	250
2	P	0.842	171	231	2	C	0.498	166	224
3	C	1.010	34	45	3	C	0.756	171	230
4	P	0.576	168	227	4	C	0.150	182	245
5	P	1.885	98	132	5	P	0.387	179	239
6	C	0.357	163	220	6	P	0.755	165	222
7	P	0.495	171	228	7	C	0.421	183	247
8	C	2.448	0	0	8	P	0.737	184	248
9	C	1.261	102	132	9	C	0.280	174	232

Tabelis 4 on välja toodud ülevaade igast filtrist. Filtrite indekseerimine algab nullist. Esimesena leiti, millist klassi filter tuvastab. Juhul, kui vastava filtri lineaarse kihi kaal on positiivne, tuvastatakse kontrollgrupi isikut (C) ning vastasel juhul patsienti (P). Läved arvutati üleüldise puhtuse väärtusega 0,75. Välja on toodud veerus „õigeid“, mitu n-grammi olid lõpliku otsustuse osas informatiivsed ehk kui filtri klass ühtis terve võrgu klassifikatsiooniotsusega. Veerus „kokku“ on välja toodud kõigi filtri läve ületanud n-grammide arv. Nagu ka eelmine joonis kinnitab, on osad filtrid ebainformatiivsed. Need filtrid ei pruugi anda niivõrd olulist informatsiooni klassifikatsiooniotsuse jaoks.

Informatiivsete ja ebainformatiivsete n-grammide analüüs töötas enamjaolt hästi. Leidus ka n-gramme, mis filtri läve ei ületanud, kuid mis intuitiivselt peale vaadates tundusid klassifikatsiooniotsuse mõttes väga olulised olema. Lävede valikusse peaks suhtuma kriitiliselt. Jacovi et al. artiklis (2018) väljapakutud lävede leidmise algoritm ei pruugi olla optimaalne. Siiski, lävede hüpotees ka DementiaBanki andmestikul andis teatud mõttes pigem häid tulemusi, sest testimistäpsus tõusis kõige optimaalsema puhtuse väärtuse juures.

5.2 Sõnede aktivatsioonivektorid

Selles alapeatükis vaadeldakse tulemusi neljast aspektist: kõrgeima aktivatsiooniga n-grammid filtrite kaupa, kõrgeima aktivatsiooniga sõned filtrite kaupa, kõrgeima aktivatsiooniga n-grammid kõikide filtrite koha ning kõrgeima aktivatsiooniga sõned kõikide filtrite kohta.

Välja pole valitud ainult läve ületavad n-grammid – vaadeldakse kõiki n-gramme. Isegi, kui mingis filtris ükski n-gramm läve ei ületanud, annab kõrgeima aktivatsiooniga n-grammide vaatlemine väärtuslikku informatsiooni.

Tabel 5. Kõrgeima aktivatsiooniga n-grammid filtrite kaupa koos sõnede aktivatsioonidega.

Filter	N-gramm	Skoor	Sõnede aktivatsioonid			
0	floor	3.04	3.04			
1	space	3.56	3.56			
2	some	3.14	3.14			
3	by	2.28	2.28			
4	huh	3.41	3.41			
5	open	3.44	3.44			
6	helping	5.98	5.98			
7	lip	4.00	4.00			
8	bench	1.04	1.04			
9	chair	3.48	3.48			
0	went overboard	5.10	3.33	1.77		
1	some plates	4.35	3.07	1.28		
2	shh quiet	5.04	4.35	0.69		
3	kitchen cabinets	2.82	1.64	1.18		
4	spilled ,	4.16	2.98	1.18		
5	my !	4.08	2.73	1.35		
6	living room	3.38	2.00	1.38		
7	help and	4.24	3.89	0.35		
8	the stool	2.48	0.67	1.81		
9	wind is	3.46	2.29	1.17		
0	hm hm cup	3.88	1.28	1.97	0.63	
1	towards her mouth	4.83	1.69	1.13	2.01	
2	's kinda sloppy	3.72	1.38	1.03	1.31	
3	stool is tipping	5.16	0.97	0.60	3.59	
4	. um mother	2.92	-0.13	0.74	2.31	
5	birds , geese	3.51	1.76	0.38	1.37	
6	bottom ? mama	5.38	2.57	0.59	2.22	
7	towards her mouth	3.81	2.44	-0.02	1.39	
8	saying “ shh	6.11	1.88	1.36	2.87	
9	finger pointed sort	3.02	1.86	1.15	0.01	
0	pan uh pan .	4.07	0.90	0.37	1.41	1.39
1	up towards her mouth	5.23	0.22	1.31	1.07	2.63
2	. uh the lid	3.87	-0.60	1.38	0.17	2.92
3	water overflowing . um	5.62	0.58	3.37	-0.19	1.86
4	feet are getting wet	5.22	1.39	2.35	-0.20	1.68
5	, huh ? looking	4.86	0.67	2.96	0.87	0.36
6	uh ... ah !	5.63	1.82	1.86	0.33	1.62
7	water is overflowing onto	6.38	1.33	1.45	3.53	0.07
8	climb the uh ...	3.71	0.00	0.68	0.90	2.13
9	um the children are	3.91	1.11	0.05	1.86	0.89

Tabelis 5 on näha, et filtrid õpivad tuvastama erineva koguaktivatsiooniga n-gramme. Veerg „skoor“ tähistab üksikute alamsõnede aktivatsioonide summat. Osad filtrid on suutelised ära õppima palju kõrgema aktivatsiooniga n-gramme kui teised filtrid. Näiteks unigrammide kuues filter sai kõige suuremaks aktivatsiooniks 5,98, kuid kaheksas filter vaid 1,04. Ka pikemate n-grammide aktivatsioonid on üleüldiselt suuremad, sest liitma peab rohkem üksikute sõnede aktivatsioone. Siiski, ei ole aktivatsioonid märkimisväärselt suuremad, sest näiteks neligrammide hulgas leidub ohkamisi, artikleid, eessõnu, määrsõnu ja kirjavahemärke, mille aktivatsioonid pole niivõrd kõrged kui teistel sama n-grammi sõnedel. Selliste sõnede hulka kuuluvad näiteks „the“, „ah“, „her“ ja „up“.

Tabel 6. Kõrgeimate aktivatsioonidega bi-, tri- ja neligrammide sõned filtrite kaupa koos koguaktivatsiooniga.

Filter	1. sõne		2. sõne		3. sõne		4. sõne		Skoor
0	went	3.33	eaten	3.17					6.50
1	some	3.07	mama	1.98					5.05
2	shh	4.35	hm	2.61					6.96
3	kitchen	1.64	cabinets	1.18					2.82
4	spilled	2.98	,	1.18					4.16
5	my	2.73	percent	2.06					4.79
6	people	2.61	billowing	1.93					4.54
7	venture	3.89	salient	1.97					5.86
8	the	0.67	stool	1.81					2.48
9	wind	2.29	is	1.17					3.46
0	hm	1.28	hm	1.97	flooding	1.83			5.08
1	raising	1.75	open	1.44	mouth	2.01			5.20
2	lady	1.78	hm	1.45	spilled	2.51			5.74
3	overflowing	2.13	condition	1.23	tipping	3.59			6.95
4	whew	1.04	towards	1.05	mother	2.31			4.40
5	towards	2.24	shrub	1.00	stool	1.84			5.08
6	bottom	2.57	forgot	1.40	pan	2.27			6.24
7	towards	2.44	uh	0.82	stool	2.18			5.44
8	mommy	2.03	“	1.36	mama	2.93			6.32
9	arm	2.27	groomed	2.30	weeds	1.05			5.62
0	tair	1.30	hm	1.08	huh	2.17	dinner	2.19	6.74
1	nine	0.94	towards	1.31	tea	1.50	mouth	2.63	6.38
2	dishes	1.29	uh	1.38	towards	2.23	lid	2.92	7.82
3	coffee	2.21	overflowing	3.37	expression	1.23	um	1.86	8.67
4	wind	1.79	are	2.35	slower	1.96	wet	1.68	7.78
5	p	2.60	huh	2.96	dear	2.74	landscaping	1.81	10.11
6	uh	1.82	”	1.94	huh	1.18	jars	1.97	6.91

7	blowing	2.22	shrub	1.92	overflowing	3.53	expression	0.94	8.61
8	begging	1.96	interested	1.02	ah	1.31	salient	3.04	7.33
9	um	1.11	towards	1.38	children	1.86	upside	1.62	5.97

Tabelis 6 on välja toodud iga filtri kohta, millised n-grammi alamsõned on kõige kõrgema aktivatsiooniga. Vaadates iga filtri poolt maksimeeritud alamsõne ühendatuna, saame väga kunstlikuna näiva n-grammi. Näiteks neligrammide viienda filtri poolt maksimeeritud sõnede ühend oleks „p huh dear landscaping“, mis ei ole loomulikus tekstis esinev neligramm. Vaadates uuesti tabelit 5 ning võrrelda selle tulemust tabeliga 6, siis võib märgata, et loomulikus kõnes esinevates n-grammides on vähemalt üks sõne, mille aktivatsioon on tunduvalt madalam kui vastaval kohal olev maksimaalse aktivatsiooniga sõne. Näiteks neligrammide seitsmenda filtri poolt on saanud maksimaalse kolmanda alamsõne aktivatsiooni n-gramm „water is overflowing onto“ (aktivatsioonid 1.33; 1.45; 3.53; 0.07), kuid selle filtri iga alamsõne aktivatsiooni maksimeerides oleks saanud neligrammi „blowing shrub overflowing expression“ (2.22; 1.92; 3.53; 0.94). Täpselt sama väitis ka Jacovi et al. oma artiklis (2018). Protsentuaalne erinevus üle kõikide filtrite kõige kõrgema loomulikus kõnes esineva n-grammi aktivatsiooni ja kõige kõrgema kunstlikult moodustatud n-grammi aktivatsiooni vahel oli 25%, mis oli natukene madalam kui Jacovi et al. saadud tulemus, 30%.

Tabel 7. Kõige kõrgema aktivatsiooniga 10 n-grammi.

Koht	N-gramm	Filter	Skoor	Sõnede aktivatsioonid	
1	helping	6	5.98	5.98	
2	help	6	5.07	5.07	
3	poor	6	4.10	4.10	
4	worst	6	4.01	4.01	
5	lip	7	4.00	4.00	
6	space	1	3.56	3.56	
7	arm	7	3.50	3.50	
8	chair	9	3.48	3.48	
9	open	5	3.44	3.44	
10	huh	4	3.41	3.41	
1	went overboard	0	5.10	3.33	1.77
2	shh quiet	2	5.04	4.35	0.69
3	pearl mom	2	4.88	3.98	0.90
4	hm hm	2	4.85	2.24	2.61
5	shh to	2	4.69	4.35	0.34
6	plates waiting	2	4.58	3.77	0.81
7	shh while	2	4.51	4.35	0.16
8	some plates	1	4.35	3.07	1.28

9	shh .	2	4.32	4.35	-0.03		
10	shh ”	2	4.32	4.35	-0.03		
1	saying “ shh	8	6.11	1.88	1.36	2.87	
2	bottom ? mama	6	5.38	2.57	0.59	2.22	
3	stool is tipping	3	5.16	0.97	0.60	3.59	
4	towards her mouth	1	4.83	1.69	1.13	2.01	
5	here 's mama	8	4.81	0.57	1.31	2.93	
6	here poor mama	8	4.77	0.57	1.27	2.93	
7	dishsink is overflowing	3	4.73	1.11	0.60	3.02	
8	standing is tipping	3	4.61	0.42	0.60	3.59	
9	whispering or motioning	3	4.57	1.18	0.17	3.22	
10	water is overflowing	3	4.44	0.82	0.60	3.02	
1	water is overflowing onto	7	6.38	1.33	1.45	3.53	0.07
2	water is overflowing .	7	6.10	1.33	1.45	3.53	-0.21
3	water is overflowing in	7	5.68	1.33	1.45	3.53	-0.63
4	uh ... ah !	6	5.63	1.82	1.86	0.33	1.62
5	water overflowing . um	3	5.62	0.58	3.37	-0.19	1.86
6	lid is leaning against	7	5.45	1.05	1.45	2.09	0.86
7	water is overflowing the	7	5.27	1.33	1.45	3.53	-1.04
8	feet are getting wet	4	5.23	1.39	2.35	-0.20	1.69
9	up towards her mouth	1	5.22	0.21	1.31	1.07	2.63
10	finger towards her mouth	1	5.21	0.21	1.30	1.07	2.63

Tabelis 7 on välja toodud kõige kõrgemate aktivatsioonidega n-grammid, kui arvestada kõiki filtreid. See aitab iseloomustada, millised n-grammid aitavad teksti erinevatesse klassidesse kõige suuremal määral eristada. Unigrammide hulgas on märgata, et kuues filter seab unigrammidele vastavusse küllaltki kõrged aktivatsioonid – võib oletada, et mudel peab unigrammide kuuendat filtrit eriti oluliseks. Bigrammide hulgas domineerib teine filter, mis tuvastab patsiendi klassi. Esineb ohtralt bigramme, mis sisaldavad sõne „shh“. Tundub ka mõistetav, et patsientidel esineb rohkem lühikesi ohkeid kui kontrollgrupil. Trigrammide hulgas domineerivad filtrid kolm ja kaheksa. Neligrammide seas on kõige olulisem filter seitsmes. Väga oluliseks peetakse vee ülevoolamise mainimist, kasutades sõne „overflowing“ – seda sõne sisaldavad n-grammid saavad üldiselt suhteliselt kõrge aktivatsiooni.

Tabel 8. Kõige kõrgema aktivatsiooniga 10 sõne.

	1. sõne			2. sõne			3. sõne			4. sõne		
Koht	Sõne	Filter	Skoor									
1	shh	2	4.35	eaten	0	3.17						
2	pearl	2	3.98	hm	2	2.61						
3	venture	7	3.89	mama	0	2.24						
4	help	7	3.89	percent	5	2.06						
5	plates	2	3.77	tin	5	1.99						
6	whew	2	3.70	mama	1	1.98						
7	mama	2	3.68	windows	5	1.98						
8	cake	2	3.52	salient	7	1.97						
9	dinner	2	3.44	billowing	6	1.93						
10	went	0	3.33	jars	1	1.93						
1	bottom	6	2.57	groomed	9	2.30	tipping	3	3.59			
2	towards	7	2.44	hm	0	1.97	motioning	3	3.22			
3	reach	6	2.30	uh	0	1.55	overflowing	3	3.02			
4	arm	9	2.27	hek	0	1.55	mama	8	2.93			
5	towards	5	2.24	hm	2	1.45	shh	8	2.87			
6	overflowing	3	2.13	open	1	1.44	overflow	3	2.72			
7	mommy	8	2.03	huh	0	1.42	um	3	2.71			
8	saying	8	1.88	forgot	6	1.40	drooping	3	2.65			
9	leg	6	1.86	“	8	1.36	spilled	2	2.51			
10	finger	9	1.86	his	8	1.32	crawling	3	2.37			
1	p	5	2.60	overflowing	3	3.37	overflowing	7	3.53	salient	8	3.04
2	cookies	5	2.49	overflow	3	3.05	overflow	7	3.18	lid	2	2.92
3	blowing	7	2.22	huh	5	2.96	nondescript	7	2.80	mouth	1	2.63
4	coffee	3	2.21	wet	3	2.61	dear	5	2.74	nose	1	2.45
5	garage	3	2.00	hazardous	3	2.54	drooping	7	2.68	arm	1	2.24
6	begging	8	1.96	washed	3	2.45	motioning	7	2.65	faucet	2	2.23
7	alright	8	1.94	are	4	2.35	jars	5	2.55	lip	1	2.22
8	boing	8	1.87	shorts	3	2.22	please	5	2.42	dinner	0	2.19
9	uh	6	1.82	nondescript	3	2.09	yeah	5	2.36	feet	1	2.19
10	kitchen	3	1.80	overflows	3	2.03	towards	2	2.23	knees	1	2.16

Tabelis 8 on välja toodud kümme kõige suurema aktivatsiooniga sõne kõikide filtrite kohta iga n-grammi alamsõne järgi. Bigrammide esimene sõne saavutas tunduvalt kõrgemad aktivatsioonid kui teine sõne. Võib oletada, et bigrammide filtrid seadsid esimesele sõnele suuremad kaalud. Trigrammide kolmas sõne oli seevastu tunduvalt kõrgema aktivatsiooniga kui esimene ja teine. Kolmandat sõne maksimeerib kolmas filter, mis enamjaolt tuvastab keerukamaid sõnu nagu näiteks „tipping“, „motioning“, „overflowing“, „drooping“ ning „crawling“. Keerukamad ja harvaesinevamad sõned saavad omale kõrgema aktivatsiooni.

Alapeatükis toodi välja konkreetsed andmed: milliseid n-gramme peab iga filter kõige olulisemaks, milliseid alamsõnesid peab iga filter kõige olulisemaks, milliseid n-gramme peab kogu neurovõrk kõige olulisemaks ning milliseid alamsõnesid peab kogu neurovõrk kõige olulisemaks. See info aitab mõista, millised n-grammid on teksti klassifitseerimise seisukohast kõige olulisemad. Samuti veenduti, et alamsõnede aktivatsioone ei maksimeerita – neil on oma kindel muster. Loomulikus keeles esinevate n-grammide korral leidub tavaliselt vähemalt üks alamsõne, mille aktivatsioon on tunduvalt väiksem kui teiste alamsõnede aktivatsioonid, täpselt nagu leidis ka Jacovi et al. (2018).

5.3 Aktivatsioonivektorite klasterdamine

Selles alapeatükis klasterdatakse bi-, tri- ja neligrammid, kasutades algoritmi *mean shift clustering*. Klasterdamist tehakse neljal viisil: kõikide n-grammide klasterdamine, kordusteta n-grammide klasterdamine, klasterdamine koos osade punktide väljajätmisega ning ainult filtri läve ületavate n-grammide klasterdamine. Klasterdamist tehti teegi scikit-learn meetodiga `MeanShift`[7], kasutades vaikeparameetreid (*bandwidth* jäeti määramata).

Esmalt klasterdati kõik n-grammid vastavalt alamsõnede aktivatsioonidele ning kõik aktivatsioonivektorid jäeti vaatlusse sisse. Klasterduse tulemus ei olnud hea. Klastreid tekkis hulgaliselt. Näiteks bigrammide teise ja viienda filtri n-grammid jaotati kaheksasse erinevasse klastrisse. Suurem osa tekkinud klastritest sisaldasid endas väga väikese osa kõikidest bigrammidest. Leidus üks suurem klaster, kuhu enamjaolt paigutusid kõik bigrammid – neisse klastritesse klasterdus üle 90% kõikidest n-grammidest.

Tabel 9. Neligrammide neljanda filtri klasterdamise tulemus.

Neligrammide neljas filter						
Klaster	Isendeid	Osakaal	Tsentroid			
1	47665	97.48%	-0.51	-0.19	-1.21	-0.50
2	500	1.02%	-0.47	1.90	-0.90	-0.49
3	162	0.33%	0.46	1.86	-0.09	-0.40
4	277	0.57%	0.75	-0.27	1.84	-0.03
5	67	0.14%	0.61	1.76	1.59	-0.10
6	9	0.02%	1.39	2.35	-0.20	1.50
7	143	0.29%	1.79	-0.98	0.41	-0.06
8	28	0.06%	-0.94	1.75	-0.13	1.47
9	48	0.10%	1.39	-0.98	-0.20	1.68

[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

Tri- ja neligrammide tulemus oli väga sarnane – jällegi tekkis üks väga suur klaster ja palju väikeseid klastreid, millesse halvimal juhul paigutus ainult üks n-gramm. Tabelis 9 on näha, et neligrammide neljanda filtri n-grammid jaotati üheksasse erinevasse klastrisse. Klastris 1 on enamus n-gramme, üle 97%.

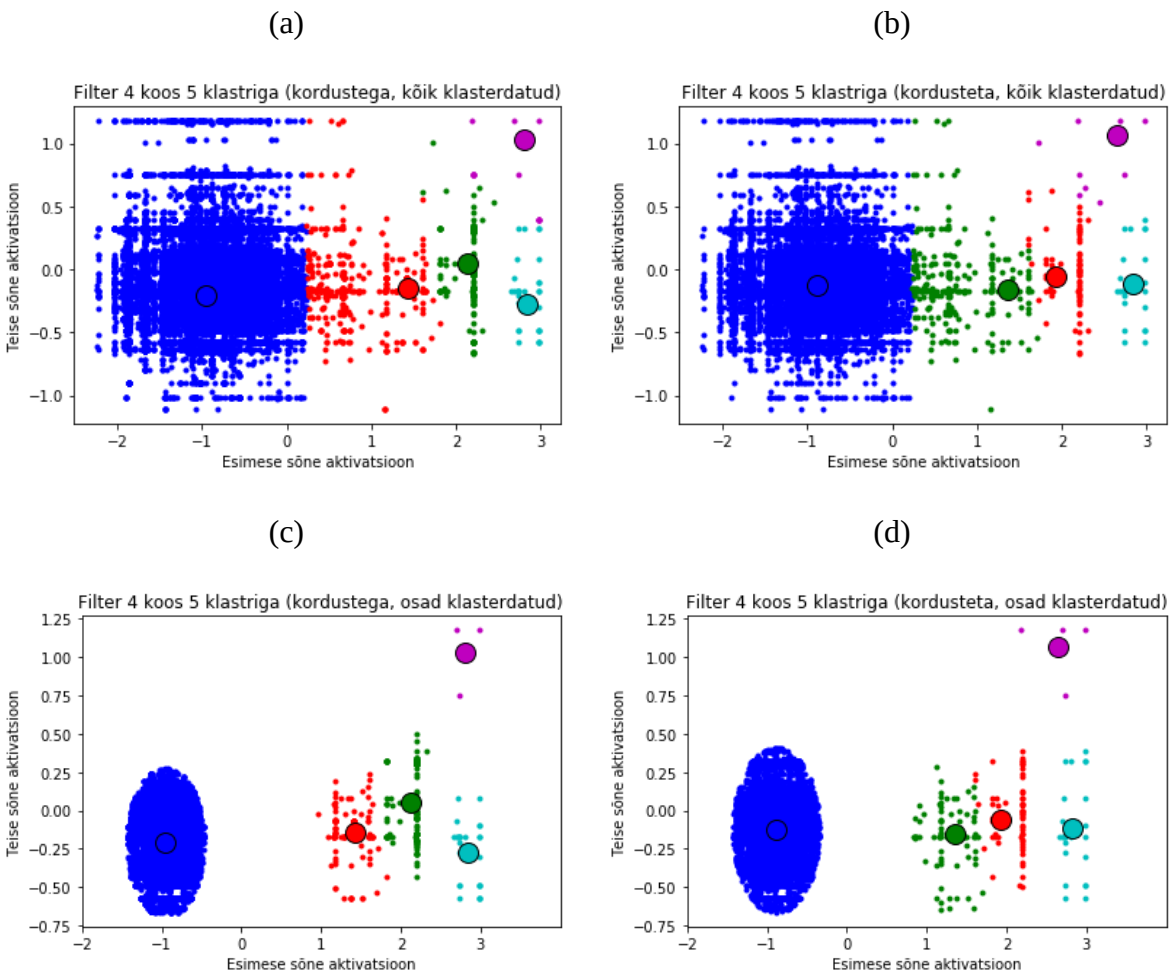
Tulemuste paremaks tõlgendamiseks on bigrammide klasterdamise tulemus esitatud joonisena. X-teljel on esimese alamsõne aktivatsioon ning y-teljel on teise alamsõne aktivatsioon ning kõik aktivatsioonivektorite paarid on paigutatud joonisele.

Joonisel 6 (a) on esitatud bigrammide neljanda filtri klasterduse tulemus, kus sisalduvad kõik bigrammid ning ühtegi neist ei ole ka välja jäetud. Jooniselt on keeruline eristada erinevaid klastreid (kui jätta klastrite värvid vaatluse alt välja), praegune klasterdamine näib kunstlik.

Üheks võimaluseks klasterdamist parandada, oleks korduste väljajätmine. Näiteks kui erinevates tekstides sisaldub üks ja sama bigramm rohkem kui üks kord, siis arvestatakse klasterdamisel seda ühekordselt. Samuti juhtudel, kui ühes tekstis on üks ja sama n-gramm mitmekordselt, siis arvestatakse joonisel vektorit ühekordselt. Joonisel 6 (b) on nähtav sellise klasterduse tulemus. Saadud joonis näeb välja väga sarnane joonisega (a), peaaegu identne. Teiste filtrite korral koostati mõningatel juhtudel vähem klastreid.

Teegi scikit-learn klasterdamise meetodi `MeanShift` üheks parameetriks on ka `cluster_all`. Kui see lülitada välja, siis punktid, mis on klastri tsentroidist piisavalt kaugel, jäetakse vaatluse alt välja. Kõikide aktivatsioonivektorite osalist klasterdamist on näha joonisel 6 (c). Nüüd tunduks, et klasterdamine annab palju parema tulemuse, kuigi mulje on petlik, sest väga paljud väärtused on jäetud kõrvale.

Joonisel 6 (d) on välja toodud klasterdamise tulemus, kui arvestatud on kõiki aktivatsioonivektoreid ainult ühekordselt ning kõik punktid ei ole klasterdatud. Kaks viimast joonist on omavahel väga sarnased, leidub üksikuid erinevusi.



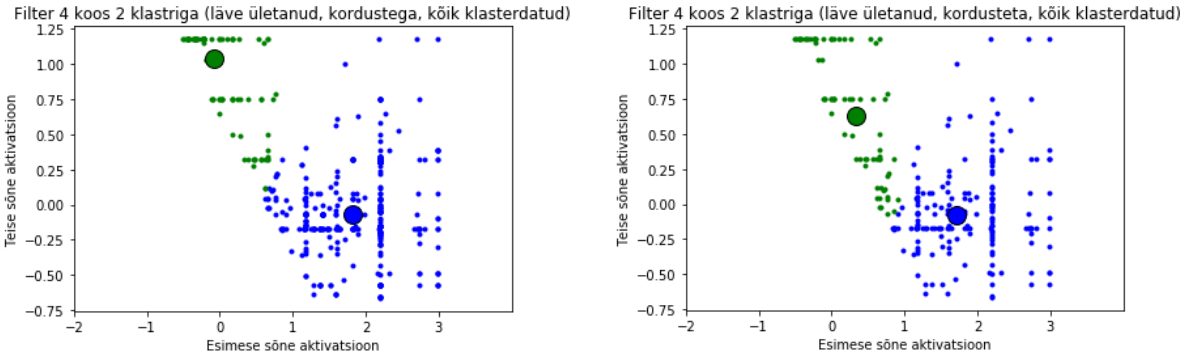
Joonis 6. Kõikide bigrammide klasterdamise tulemus.

Järgnevalt proovitakse sooritada klasterdamist ainult nende n-grammide aktivatsioonivektoritega, mis ületavad vastava filtri läve. Filtri läve ületavate n-grammide klasterdamisel saadud tulemus oli kordades parem. Tuleb selgelt välja, et filtrid tuvastavad erinevaid aktivatsioonimustreid, täpselt nagu Jacovi et al. artiklis (2018) oli ka kirjeldatud.

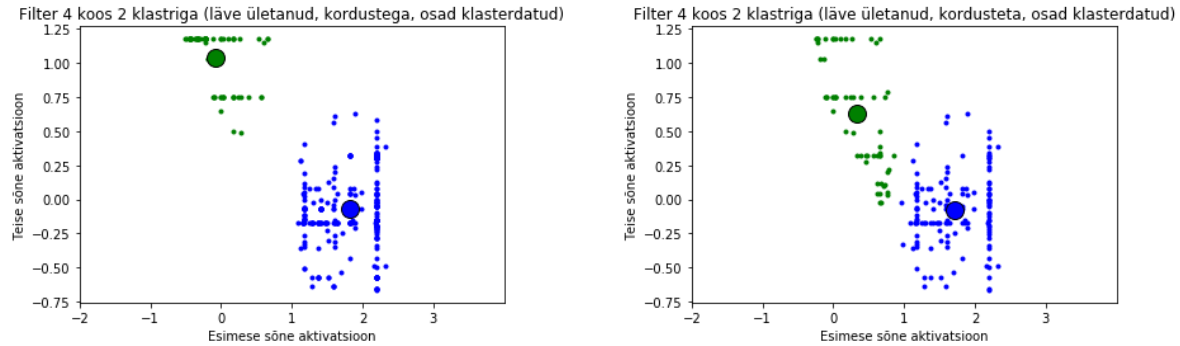
Vaatleme jällegi parema võrdluse nimel bigrammide neljandat filtrit, mis on esitatud joonisel 7. Klasterdus on sooritatud täpselt sama moodi nagu joonisel 6, ainuke erinevus on see, et klasterdus on tehtud peale läve mitte ületavate bigrammide väljafiltreerimist. Selgelt tuleb esile, et tekib kaks suuremat klastrit, mis on üksteisest suhteliselt hästi eraldatud. Märgata on ka graafiku teatud kohtades punktidest moodustunud vertikaalseid ja horisontaalseid jooni. Järelikult leidub arvukalt bigramme, mille esimene või teine sõne on samad ning seesama filter seab neile n-grammidele kõrged aktivatsioonid. On kaks põhjust, miks selline nähtus esineb. Esiteks, ühes ja samas tekstis võib leiduda ühte ja sama n-grammi mitu korda. Teiseks, mitmetes erinevates tekstides võib sisalduda üks ja sama n-gramm. Nähtavasti

needsamad n-grammid ületavad vastavate filtrite läve ka teiste tekstide puhul. Siit võib järeldada, et n-grammides on olulised alamsõned, mis kõrgendavad n-grammi koguaktivatsiooni, nii et see ületaks ka filtri läve.

(a) (b)



(c) (d)



Joonis 7. Läve ületavate bigrammide klasterdamise tulemus.

Järgmisena vaadeldakse konkreetseid trigramme – millistesse klastritesse need jaotusid ning mis on nendel klastritel ühist.

Tabel 10. Viienda filtri kuus kõige kõrgema aktivatsiooniga mõlema klastri trigrammi.

Trigramm	Summa	1. sõne	2. sõne	3. sõne	Klaster
tsentroid	1.44	-0.31	-0.04	1.79	1
climbing a stool	2.90	0.90	0.16	1.84	1
off a stool	2.59	0.59	0.16	1.84	1
tipping the stool	2.55	0.67	0.05	1.84	1
off the stool	2.48	0.59	0.05	1.84	1
off that stool	2.02	0.59	-0.41	1.84	1
upsetting the stool	1.97	0.09	0.05	1.84	1

tsentroid	1.22	0.28	0.36	0.59	2
birds , geese	3.51	1.76	0.38	1.37	2
slipping from stool	3.32	1.35	0.14	1.84	2
towards her mouth	3.29	2.24	-0.22	1.27	2
moving , I	1.73	1.55	0.38	-0.20	2
moving her finger	1.72	1.55	-0.22	0.39	2
flowers or weeds	1.71	0.73	0.11	0.87	2

Tabelis 10 on toodud välja viienda filtri kuus kõige kõrgema aktivatsiooniga trigrammi mõlema klastri kohta. Filtri läve ületas kokku 61 n-grammi. Neist 34 paigutus esimesse klastrisse (55,74%) ning 27 teise (44,26%). Filtris paigutusid trigrammid klastritesse peaaegu võrdselt ning klastreid tekkis ainult kaks tükki, seepärast valiti ka vastav filter analüüsiks välja. Täpselt nagu Jacovi et al. artikliski (2018), pole see filter tingimata homogeenne – teise klastrisse sattunud trigrammid näivad suhteliselt juhuslikud. Küll aga esimese klastri trigrammid sisaldavad endas sõna „tool“. Võib järeldada, et ka DementiaBanki andmestiku korral võivad filtrid tuvastada semantiliselt sarnaseid n-gramme, kuid ka oluline osa on sõnede aktivatsioonimustril, mis määrab, milliseid n-gramme antud filter tuvastab.

Esialgsel, kõikide n-grammide aktivatsioonivektorite klasterdamisel saadi halb tulemus: ei moodustunud eriti hästieristuvaid klastreid ning klastreid tekkis ohtralt. Samuti suurem osa klastritest sisaldas endas väga väikese koguse aktivatsioonivektoreid. Seejärel prooviti klasterdada ainult läve ületanud n-grammide aktivatsioonivektoreid. Klasterduse tulemus muutus oluliselt paremaks. Üldiselt tekkis vähem klastreid ning enam polnud klastrid väga väikesed. Nähtus annab lisakinnitust informatiivsete ja ebainformatiivsete n-grammide hüpoteesi paikapidavusele (vaata peatükist 5.1).

Bakalaureusetöö autoril tekkis peale analüüsi teostamist mõte, et filtrite lävesid võiks leida teistmoodi. Nägime, et igal filtril on tekkinud oma klastrid, igaühel neist on tsentroid ehk kõige iseloomulikum aktivatsioonimuster. Hetkel määratakse filtri läve universaalselt kõikide n-grammide koguaktivatsioonide põhjal. Kui teha eelnevalt selgeks, millisesse klastrisse n-gramm kuulub, siis peaks olema võimalik määrata lävi igale klastrile eraldi. Seeläbi otsustatakse n-grammi informatiivsuse üle, vaadates, kas n-grammi koguaktivatsioon ületab konkreetse klastri läve või mitte. Selline lähenemine võib anda veelgi parema tulemuse, sest analüüsi kaasatakse rohkem informatiivseid n-gramme.

5.4 Vastandlikud n-grammid

Selles alapeatükis vaadeldakse vastandlikke n-gramme. Valitakse välja üks konkreetne n-gramm ning kõik muud n-grammid, mille Hammingi kaugus on esialgsest n-grammist täpselt üks.

Tabel 11. Tri- ja neligrammide vastandlikud n-grammid koos aktivatsioonide erinevustega.

N-gramm	1. sõne	2. sõne	3. sõne	4. sõne	Aktivatsioon	Erinevus
stool is tipping	0.97	0.60	3.59		5.15	0.00
stool is coming			-2.16		-0.59	5.75
stool is going			-1.83		-0.26	5.42
stool is turning			1.01		2.58	2.58
stool is tilted			1.08		2.65	2.51
which is tipping	-0.97				3.22	1.94
socks are drooping	1.37	0.14	2.65		4.16	0.00
socks are alright			-0.83		0.68	3.48
water is overflowing .	1.33	1.45	3.53	-0.21	6.11	0.00
water is flowing .			-0.67		1.90	4.20
water is running .			0.04		2.61	3.49
water 's overflowing .		-1.81			2.84	3.27
water is splashing .			0.49		3.06	3.04

Tabelis 11 on esitatud andmed n-grammidest ja toodud välja kõige kõrgema aktivatsiooniga vastandlikud n-grammid. Välja on toodud ka erinevus: kui palju sõne asendamise tulemusena saadud n-gramm on madalama aktivatsiooniga. Esimese näitena on valitud n-grammiks „stool is tipping“ ehk tool hakkab ümber kukkuma. Vastandlikud n-grammid on mudeli meelest „stool is coming“, „stool is going“ ja teised. Need konstruktsioonid ei ole aga loomulikule terve inimese kõnele omased. Mudel püüab nendele sõnedele madala aktivatsiooni andmisega mõjutada klassifikatsiooniotsust vastupidises suunas. Üks huvitavamaid näiteid oli sokkide kirjeldamine, millele kontrollgruppi tuvastav trigrammide kolmas filter oluliseks pidas. Kui kontrollgrupi isikud ütlevad, et sokid on alla vajumas, siis patsiendid väidavad, et sokkidega on kõik korras – nad ei märka seda pisidetaili.

Häid näiteid vastandlikest n-grammidest oli keerukas leida. Siiski, DementiaBanki andmestik on üpriski teistsugune põhiartikli (Jacovi, et al., 2018) andmestikust – seal teostati analüüsi filmi- ja tootearvustustele. Sellistes andmestikes on ohtralt eitusi: „mulle ei meeldi see toode“. Samuti on sellistele andmestikele omane vastandlike sõnade (antonüümide) rohkus

näiteks „kasulik“ ja „piiratud“ või „rahulolev“ ja „pettunud“. DementiaBanki andmestikus on aga inimestel palutud kirjeldada fotot. Esineb palju vähem eitusi – ei kirjeldata, mida fotol ei ole, vaid just kirjeldatakse (väga väheste eitustega), mida fotol nähakse. Seega võib järeldada, et vastandlike n-grammide analüüs on pigem sobilik meelsusanalüüsile kui DementiaBanki andmestikule. Mõningaid tulemusi saadi, kuid võrreldes eelnevate analüüsidega, olid tulemused halvemad.

5.5 Konkreetsete tekstide interpreteerimine

Selles alapeatükis tuuakse välja konkreetseid näiteid tekstidest. Vaadatakse täpselt järele, millised n-grammid valiti iga filtri poolt välja peale *max-poolingut*, mis klassi filtrid tuvastavad ning mis on filtrite läved. Eraldi on tähistatud paksu kirjaga n-grammid, mille aktivatsioonid (millele on juurde liidetud ka vabaliikmed) ületavad vastava filtri läve. Samuti esitatakse ka koguaktivatsioon ning sõnede aktivatsioonivektorid.

Näidetena tuuakse välja neli juhtu: mudeli ennustus ühtib märgendiga – mõlemad on patsiendid (tõepositiivne) ning mõlemad on kontrollgruppi kuuluvad (tõenegatiivne). Samuti vaadeldakse juhtumeid, kus mudeli ennustus pole korrektne – mudel ennustab, et tegemist on kontrollgrupi isikuga, kuid tegelikult on patsient (valenegatiivne) ning ka juhtumit, kus mudeli klassifikatsiooni kohaselt on tegemist patsiendiga, kuid tegelikult on isik kontrollgrupist (valepositiivne).

Esmalt vaadeldakse tõepositiivset näidet, kus mudeli ennustus (patsient) ühtib märgendiga (patsient).

Tabel 12. Alzheimeri patsiendi õigesti märgendatud tekst (tõepositiivne).

girl washing dishes . I see that . uh what the boy 's putting up the cookie jar or getting cookies out of the cookie jar . and his little sister is begging him . oh hurry up . come on . and he 's got the foot stool . and his her big sister is washing the dishes . that 's the big sister or the mama I do n't know which that is is drying the dishes , putting them away . she 's got the spigot on and the water 's running . and there 's her cup . she has n't them put up yet . and he 's up on the foot stool . and he 's got the cookie jars . he 's getting the cookie jars . she 's begging him for some . begging him for some cookies . she 's washing the dishes uh she 's drying the dishes . and she 's got the water on in the sink . and there 's her cup sitting down there . , to be washed . no she has an apron on and her shoes . and the drapes are pulled back in one of the rooms . see ? P								
Filter	Lävi	Klass	N-gramm	Aktivatsioon	1. sõne	2. sõne	3. sõne	4. sõne
0	3.04	P	begging	1.00	1.00			
1	0.92	C	putting	0.29	0.29			
2	3.18	P	some	3.14	3.14			

3	2.24	P	some	0.62	0.62			
4	0.96	P	**jars**	1.85	1.85			
5	0.34	C	I	0.04	0.04			
6	1.17	P	mama	1.00	1.00			
7	0.81	C	foot	0.10	0.10			
8	1.07	C	stool	0.82	0.82			
9	2.32	P	**mama**	2.85	2.85			
0	0.92	P	**got the**	1.83	2.05	-0.22		
1	1.69	P	**some cookies**	2.85	3.07	-0.22		
2	0.97	P	**mama I**	3.44	3.68	-0.24		
3	1.01	C	dishes uh	0.93	0.02	0.91		
4	0.64	P	jars .	0.06	0.22	-0.17		
5	1.98	P	her big	1.94	1.65	0.29		
6	0.36	C	be washed	0.26	0.85	-0.59		
7	0.50	P	rooms .	0.31	0.00	0.31		
8	2.45	C	foot stool	1.04	-0.78	1.81		
9	1.26	C	**drapes are**	1.36	0.41	0.95		
0	1.31	P	's got the	0.53	0.27	-0.08	0.35	
1	0.98	C	on the foot	0.54	0.17	-0.17	0.53	
2	1.29	P	's got the	0.61	1.38	-0.67	-0.10	
3	0.39	C	sister is begging	0.07	-0.76	0.60	0.23	
4	1.68	C	drying the dishes	-0.26	0.59	-0.08	-0.77	
5	0.84	C	the foot stool	0.54	-0.94	-0.35	1.84	
6	1.71	P	cup . she	0.88	1.61	0.03	-0.76	
7	1.78	C	the foot stool	1.11	-0.84	-0.24	2.18	
8	1.82	P	**or the mama**	2.75	-0.20	0.03	2.93	
9	0.14	C	**dishes uh she**	0.21	-0.62	0.45	0.38	
0	0.33	P	**rooms . see ?**	0.92	0.00	-0.36	-0.19	1.48
1	0.35	C	**on and her shoes**	1.25	0.31	-0.68	1.07	0.56
2	0.54	C	dishes . I see	0.50	1.29	0.17	-0.14	-0.83
3	0.86	C	**be washed . no**	1.75	-0.89	2.45	-0.19	0.38
4	0.18	C	**to be washed .**	0.53	0.71	1.70	-1.94	0.06
5	0.39	P	**him . oh hurry**	0.59	-0.58	-0.51	1.69	0.00
6	0.76	P	**getting the cookie jars**	1.49	-0.30	-0.47	0.30	1.97
7	0.87	C	drapes are pulled back	-0.12	0.56	1.02	-1.04	-0.65
8	0.75	P	**got the cookie jars**	2.45	0.77	0.68	-0.77	1.76
9	0.28	C	know which that is	-0.68	-0.08	-0.66	-0.73	0.79

Mudel ennustas, et tegemist on patsiendiga tõenäosusega 0,99, mis on väga kindel otsustus (vaata tabelist 12). Uni- ja bigrammide puhul tuleb selgelt välja, et enamjaolt on valitud patsiendi klassi filtrite läve ületanud n-grammid. Tri- ja neligrammide korral on läve ületanud n-grammid jaotunud peaaegu võrdselt kahe klassi vahel. Patsiendi klassi tuvastavad n-

grammid on kõrgema koguaktivatsiooniga kui kontrollgrupi klassi tuvastavad n-grammid. Intuitiivselt teksti lugedes märkame, et isik elab pildil olevale poisile kaasa öeldes: „oh hurry up . come on .“ nagu oleks isik samas fotol kujutatavas situatsioonis. Samuti ei suuda isik vahet teha, kas pildil on suur õde või ema. Isiku kõne on lihtne, ei kasutata keerukaid lause-konstruktsioone. Esineb kõhklust ning palju lühilauseid.

Järgnevalt vaadeldakse tõenegatiivset näidet, kus mudeli ennustus (kontrollgrupp) ühtib märgendiga (kontrollgrupp).

Tabel 13. Kontrollgrupi indiviidi õigesti märgendatud tekst (tõenegatiivne).

okay . the little boy is climbing up in the getting cookies out of the cookie jar and is he 's falling . the stool 's falling over . and the little girl has her hand up that she wants a cookie and he 's trying to hand her one . and the door to the cabinet is open . and the mother is washing dishes . and the the dish water is overflowing in the sink and it 's running on to the floor and she 's standing in the water . and there 's she 's drying a plate or washing it . whatever . and there 's two cups and a plate on the counter top . and it looks like the window is open . and it has curtains pulled back at the window . and there 's a little lane going around the house towards the back I imagine with shrubs . and you can see another window . C							
Filter	Lävi	Klass	N-gramm	Aktivatsioon	1. sõne	2. sõne	3. sõne	4. sõne
0	3.04	P	**floor**	3.04	3.04			
1	0.92	C	towards	0.71	0.71			
2	3.18	P	up	2.11	2.11			
3	2.24	P	over	0.66	0.66			
4	0.96	P	imagine	0.15	0.15			
5	0.34	C	**open**	3.44	3.44			
6	1.17	P	**trying**	2.40	2.40			
7	0.81	C	**cabinet**	1.62	1.62			
8	1.07	C	stool	0.82	0.82			
9	2.32	P	climbing	0.83	0.83			
0	0.92	P	pulled back	0.50	0.43	0.06		
1	1.69	P	little lane	0.69	0.69	0.00		
2	0.97	P	door to	0.43	0.09	0.34		
3	1.01	C	shrubs .	0.87	1.48	-0.60		
4	0.64	P	**whatever .**	1.65	1.81	-0.17		
5	1.98	P	her one	1.67	1.65	0.02		
6	0.36	C	**wants a**	1.08	1.00	0.08		
7	0.50	P	**whatever .**	1.86	1.55	0.31		
8	2.45	C	**the stool**	2.49	0.67	1.81		
9	1.26	C	**water is**	1.72	0.55	1.17		
0	1.31	P	wants a cookie	0.55	0.55	-0.92	0.92	
1	0.98	C	**towards the back**	1.23	1.69	-0.17	-0.29	

2	1.29	P	's she 's	0.69	1.38	-0.86	0.17	
3	0.39	C	**water is overflo-wing**	4.44	0.82	0.60	3.02	
4	1.68	C	and the mother	0.88	-1.35	-0.08	2.31	
5	0.84	C	**towards the back**	1.20	2.24	0.05	-1.08	
6	1.71	P	has curtains pulled	0.76	-0.05	0.68	0.13	
7	1.78	C	towards the back	1.45	2.44	-0.03	-0.96	
8	1.82	P	she 's drying	1.27	0.40	1.31	-0.44	
9	0.14	C	with shrubs .	0.01	-0.25	0.86	-0.60	
0	0.33	P	jar and is he	-0.22	-0.92	-0.48	-0.12	1.30
1	0.35	C	**house towards the back**	0.96	-0.36	1.31	-0.06	0.08
2	0.54	C	**looks like the win-dow**	0.68	-0.63	-0.84	0.17	1.98
3	0.86	C	**is overflowing in the**	2.71	0.15	3.37	-1.36	0.55
4	0.18	C	**window is open .**	3.06	0.81	0.36	1.84	0.06
5	0.39	P	jar and is he	-0.56	0.59	-0.52	-0.92	0.29
6	0.76	P	. whatever . and	-0.32	-0.72	1.03	-0.31	-0.32
7	0.87	C	**water is overflo-wing in**	5.69	1.33	1.45	3.53	-0.63
8	0.75	P	cabinet is open .	0.22	0.24	-0.15	0.03	0.11
9	0.28	C	**like the window is**	0.93	-1.12	0.05	1.20	0.79

Mudel ennustas, et tegemist on kontrollgrupi indiviidiga tõenäosusega 0,99, mis on jällegi väga kindel otsustus (vaata tabelist 13). Erinevalt eelmisest tabelist (tabel 12), on läve ületanud tri- ja neligrammide filtrid ainult kontrollgrupi klassi omad. *Max-poolingu* läbinud n-grammid on oma konstruktsioonilt keerukamad. Tri- ja neligrammide hulgas pööratakse erilist tähelepanu aknale, sealt paistvale majale ning vee valamust väljavoolamisele. Isiku kõne on sujuv, kasutatakse keerukamaid väljendeid, ei kõhelda.

Järgnevalt vaadeldakse valenegatiivset näidet, kus mudeli ennustus (kontrollgrupp) ei ühti märgendiga (patsient).

Tabel 14. Patsiendi valesti märgendatud tekst (valenegatiivne).

hm . the little boy is on the stool which is tipping and he wants cookies to give to his sister . he has a handful of cookies in one hand already and the little girl is eating a cookie . the mother is washing dishes absentmindedly . she let the sink run over and it 's all on the floor . there 's a window with you can see their driveway . either bushes or trees and grass a big window . and a tree coming up here . and the part of the roof . and the cookie says cookie jar . P								
Filter	Lävi	Klass	N-gramm	Aktivatsioon	1. sõne	2. sõne	3. sõne	4. sõne
0	3.04	P	**floor**	3.04	3.04			
1	0.92	C	**already**	0.98	0.98			

2	3.18	P	all	2.17	2.17			
3	2.24	P	their	0.79	0.79			
4	0.96	P	**here**	1.23	1.23			
5	0.34	C	grass	0.13	0.13			
6	1.17	P	coming	0.62	0.62			
7	0.81	C	says	0.12	0.12			
8	1.07	C	stool	0.82	0.82			
9	2.32	P	hm	0.83	0.83			
0	0.92	P	already and	0.62	0.94	-0.31		
1	1.69	P	dishes absentmin-dedly	0.85	0.85	0.00		
2	0.97	P	**hm .**	2.22	2.24	-0.03		
3	1.01	C	window .	0.16	0.76	-0.60		
4	0.64	P	tipping and	0.21	-0.11	0.32		
5	1.98	P	his sister	0.98	1.82	-0.84		
6	0.36	C	wants cookies	0.27	1.00	-0.73		
7	0.50	P	tipping and	0.42	0.07	0.35		
8	2.45	C	**the stool**	2.49	0.67	1.81		
9	1.26	C	grass a	0.63	0.12	0.51		
0	1.31	P	girl is eating	0.95	0.75	-1.06	1.26	
1	0.98	C	**on the stool**	1.24	0.17	-0.17	1.24	
2	1.29	P	**says cookie jar**	1.35	0.36	0.31	0.68	
3	0.39	C	**which is tipping**	3.22	-0.97	0.60	3.59	
4	1.68	C	**. the mother**	2.10	-0.13	-0.08	2.31	
5	0.84	C	**on the stool**	1.80	-0.09	0.05	1.84	
6	1.71	P	cookie says cookie	1.48	-0.39	0.49	1.37	
7	1.78	C	**on the stool**	2.48	0.33	-0.03	2.18	
8	1.82	P	he wants cookies	0.67	0.72	0.14	-0.19	
9	0.14	C	**or trees and**	0.87	-0.43	1.19	0.11	
0	0.33	P	dishes absentmin-dedly . she	0.02	-0.38	0.00	-0.35	0.75
1	0.35	C	or trees and grass	0.10	-0.61	0.31	-0.23	0.63
2	0.54	C	she let the sink	-0.13	-0.43	-0.60	0.17	0.74
3	0.86	C	dishes absentmin-dedly . she	0.60	-0.16	0.00	-0.19	0.94
4	0.18	C	**which is tipping and**	0.57	-0.17	0.36	0.71	-0.32
5	0.39	P	**says cookie jar .**	1.00	-0.16	-0.52	1.50	0.17
6	0.76	P	cookie says cookie jar	0.40	-1.18	0.26	0.30	1.02
7	0.87	C	**which is tipping and**	2.73	-0.49	1.45	1.98	-0.21
8	0.75	P	run over and it	-0.08	0.06	-0.04	0.09	-0.19
9	0.28	C	**. the mother is**	1.45	-0.73	0.05	1.34	0.79

Tabelis 14 on välja toodud tulemused – mudel teeb vale otsuse. Tehakse otsus, et tegemist on kontrollgrupi isikuga tõenäosusega 0,71, mis on siiski ka kaugel täiesti kindlast otsustusest. Isiku kõne algab küllaltki sujuvalt, kasutatakse ka keerukamaid sõnu nagu näiteks „absentmindedly“. Isiku selgitus pildist lõppeb küll lausega „and the cookie says cookie jar .“, mis ei tundu küll eriti loogiline olevat ning viitab sellele, et tegemist võib olla patsiendiga. Neligrammide viies filter tuvastabki neligrammi „says cookie jar .“, mis ületab ka vastava filtri läve, kuid mille aktivatsioon ei ole väga kõrge (1,00). Võib oletada, et treeningandmestikus ei leidunud eelmainitud n-grammi eriti palju ning mudel seetõttu määras n-grammile oodatust madalama koguaktivatsiooni. Suurem osa filtri läve ületanud n-gramme kuulub kontrollgrupi klassi tuvastavate hulka. Kui vaadata teksti ning *max-poolitud* n-gramme, siis on ka mõistetav, miks mudel sellise otsustuse tegi, sest patsiendi klassile omaseid kõnekonstruktsioone on väga vähe. Samuti on treeningandmestik äärmiselt väike ning suurem andmestik aitaks mudeli täpsust parandada.

Viimasena vaadeldakse valepositiivset näidet, kus mudeli ennustus (patsient) ei ühti märgendiga (kontrollgrupp).

Tabel 15. Kontrollgrupi indiviidi valesti märgendatud tekst (valepositiivne).

uh inside the room or every place ? oh I can oh you do n't want me to memorize it ! oh . okay , the the little girl asking for the cookie from the boy who 's about to fall on his head . and she 's going I guess “ shush ” or “ give me one ” . the mother 's we do n't think she might be on drugs because uh she 's off someplace because the sink 's running over . and uh it 's summer outside because the window 's open and the grasses or the bushes look healthy ... and uh she 's drying dishes with her apron on . and the cookie jar 's looking full . that 's it . C								
Filter	Lävi	Klass	N-gramm	Aktivat-sioon	1. sõne	2. sõne	3. sõne	4. sõne
0	3.04	P	room	1.41	1.41			
1	0.92	C	**summer**	1.22	1.22			
2	3.18	P	n't	2.13	2.13			
3	2.24	P	over	0.66	0.66			
4	0.96	P	**oh**	1.63	1.63			
5	0.34	C	**open**	3.44	3.44			
6	1.17	P	**head**	2.19	2.19			
7	0.81	C	head	0.58	0.58			
8	1.07	C	memorize	0.00	0.00			
9	2.32	P	head	1.85	1.85			
0	0.92	P	**his head**	0.96	0.51	0.45		
1	1.69	P	guess “	0.46	0.52	-0.06		
2	0.97	P	**! oh**	1.28	1.60	-0.32		
3	1.01	C	and uh	0.63	-0.29	0.91		

4	0.64	P	**... and**	2.52	2.20	0.32		
5	1.98	P	her apron	1.54	1.65	-0.11		
6	0.36	C	**the room**	0.45	-0.93	1.38		
7	0.50	P	**place ?**	1.06	0.63	0.43		
8	2.45	C	the grasses	0.67	0.67	0.00		
9	1.26	C	shush ”	0.16	0.00	0.16		
0	1.31	P	girl asking for	0.60	0.75	-0.48	0.33	
1	0.98	C	with her apron	0.28	-1.08	1.13	0.23	
2	1.29	P	I can oh	0.39	-0.07	-1.34	1.80	
3	0.39	C	drying dishes with	-0.28	0.52	-0.24	-0.56	
4	1.68	C	**. the mother**	2.10	-0.13	-0.08	2.31	
5	0.84	C	bushes look healthy	-0.50	-0.14	-0.35	0.00	
6	1.71	P	place ? oh	1.11	0.65	0.59	-0.13	
7	1.78	C	off someplace because	0.00	1.33	0.00	-1.33	
8	1.82	P	**guess “ shush**	1.97	0.61	1.36	0.00	
9	0.14	C	because uh she	0.13	-0.70	0.45	0.38	
0	0.33	P	**I can oh you**	1.31	-0.03	-1.18	1.52	1.00
1	0.35	C	dishes with her apron	0.20	-1.34	-0.15	1.07	0.63
2	0.54	C	**I guess “ shush**	0.55	0.03	0.47	0.05	0.00
3	0.86	C	looking full . that	0.04	-0.41	-0.66	-0.19	1.29
4	0.18	C	**window 's open and**	1.35	0.81	-0.98	1.84	-0.32
5	0.39	P	**oh . okay ,**	1.02	1.20	-0.51	0.07	0.26
6	0.76	P	**guess “ shush ”**	1.50	0.00	1.13	0.00	0.37
7	0.87	C	bushes look healthy ...	-0.54	0.07	-0.32	0.00	-0.30
8	0.75	P	**bushes look healthy ...**	1.24	-0.90	0.01	0.00	2.13
9	0.28	C	because the window 's	-0.06	-0.55	0.05	1.20	-0.75

Tabelis 15 on välja toodud analüüsi tulemused – mudeli ennustus ei ole korrektne. Tehakse otsus, et tegemist on patsiendiga tõenäosusega 0,98, mis on täiesti vale. Filtri läve ületab rohkem patsiendi klassi tuvastavaid n-gramme ning ka mudel otsustab niimoodi. Lugedes teksti, näeme, et tekst ongi suhteliselt segane. Esineb hüüdsõnu: „oh“, „uh“ ning mõttepause „…“. Isik räägib kõhklevalt ning teksti lugedes ongi keeruline aru saada, et tegemist on kontrollgruppi kuuluva isikuga. Sellist teksti on ka inimesel keeruline klassifitseerida.

Konkreetsete tekstide interpreteerimisel vaadeldi, millised n-grammid läbisid *max-poolingu* ning lisaks vaadeldi, millised neist ületasid ka läve (märgitud paksu kirjaga). Leiti täiendavat kinnitust lävehüpoteesile – suur osa läve ületavate n-grammide filtrite klasse ühtis lõpliku ennustusega. Selle analüüsiga näeme, milliste n-grammide põhjal mudel ennustuse teeb. Samuti vaatlesime olukorda, kus mudel tegi vale ennustuse – ühte juhtu saaks parandada treeningandmete lisamisega ning teine juhtum oli niivõrd segane, et teksti ongi äärmiselt keeruline klassifitseerida.

Kokkuvõte

Bakalaureusetöö eesmärk oli rakendada artiklis (Jacovi, et al., 2018) väljapakutud interpretatsioonimeetodeid uuel andmestikul, et katsetada meetodite usaldusväärsust. Selleks programmeeriti valmis lisaks neurovõrgule artiklis kirjeldatud interpreteerimismeetodid: informatiivsete ja ebainformatiivsete n-grammide leidmine, sõnede aktivatsioonivektorite leidmine koos klasterdamisega ning vastandlike n-grammide leidmine. Samuti interpreteeriti konkreetseid tekste.

Analüüsi teostati uuel, kliinilisel andmestikul. Käesolevas uurimistöös kasutati DementiaBanki andmestikku, mis sisaldab endas Alzheimeri tõvega ja kontrollgruppi kuuluvate inimeste (transkribeeritud) kirjeldusi Bostoni küpsisevarguse fotost.

Esitati ka järgmine uurimisküsimus: kuivõrd rakendatavad on Jacovi et al. (2018) poolt välja pakutud meetodid kliinilisel andmestikul treenitud klassifitseerija interpreteerimisel? Küsimusele vastatakse erinevate analüüside haaval.

Informatiivsete ja ebainformatiivsete n-grammide analüüsis leiti, milline on optimaalseim universaalne puhtus, et mudeli täpsus oleks parim. Puhtuse põhjal leiti iga filtri lävi ning ka arvud, mitu n-grammi vastava filtri läve aktivatsiooni ületasid. Tulemused olid väga sarnased artiklis (Jacovi, et al., 2018) kirjeldatule. Leiti, et lävede valikusse peaks suhtuma kriitiliselt, sest osad informatiivsena tundunud n-grammid ei ületanud läve. Üldiselt töötas interpretatsioonimeetod hästi ning seda meetodit DementiaBanki andmestikul oli mõistlik rakendada.

Leiti ka andmestiku sõnede aktivatsioonivektorid. Toodi välja iga filtri haaval, millistel n-grammidel on kõige suuremad aktivatsioonid. Lisaks näidati ka iga n-grammi alamsõne kaupa eraldi, millised n-grammid maksimeerivad vastavaid alamsõnesid kõige rohkem. Toodi välja ka, millised n-grammid mõjutavad kogu neurovõrgu otsustust kõige rohkem ning millised n-grammid mõjutavad enim iga alamsõne kaupa neurovõrgu otsustust. Info aitab paremini mõista, millised n-grammid ning alamsõned on klassifikatsiooni seisukohast kõige olulisemad ehk milliste n-grammide või alamsõnede olemasolu kõige paremini patsiendi ja kontrollgrupi klassi üksteisest eristavad. Interpretatsioonimeetodit oli kindlasti mõistlik rakendada ka DementiaBanki andmestikul, sest erinevate (alam-)sõnede aktivatsioonides veendumine andis parema ülevaate n-grammidest, mida mudel peab oluliseks.

Seejärel klasterdati kõikide n-grammide aktivatsioonivektorid ning jõuti järeldusele, et tulemus ei ole eriti hea. Võrdlusena toodi välja ka klasterduse tulemus, kui klasterdatud on ainult filtri läve ületavate n-grammide aktivatsioonivektorid. See muutis klasterduse tulemust teatud juhtudel oluliselt selgemaks, mis annab täiendavat kinnitust filtrite läve hüpoteesi mõttekusele. Interpretatsioonimeetodit oli kindlasti mõttekas kasutada DementiaBanki andmestikuga. Meetodi kasutamine andis palju parema ülevaate erinevatest aktivatsioonimustritest, mis aitab interpreteerida neurovõrku paremal viisil.

Viimase analüüsina esitati vastandlikud n-grammid. Leiti mõningaid olemuslikult vastandlikke n-gramme, kuid tulemused ei olnud niivõrd head nagu originaalartiklis (Jacovi, et al., 2018). DementiaBanki andmestikus esineb pigem vähe eitusi, mida aga originaalartikli andmestikes oli ohtralt (filmi- ja tootearvustused). Järelikult sobib vastandlike n-grammide analüüs paremini andmestikel, mis sisaldavad rohkem eituseid. Vastandlike n-grammide analüüs oli küll DementiaBanki andmestikul rakendatav, kuid tulemus ei olnud niivõrd hea.

Kõige viimasena interpreteeriti konkreetseid tekste: toodi välja, millised n-grammid valiti välja *max-poolingu* kihist ning millised n-grammid ületasid ka filtrite läve. Analüüsiti ka juhtumeid, kus mudel tegi vale otsustuse. Mudeli ennustuste interpreteerimine aitab mõista konkreetse teksti klassifikatsiooniotsust ning võib anda selgitusi selle kohta, miks mudel mõnikord valesti klassifitseerib.

Uurimust saaks edasi arendada, püüdes muuta artiklis välja toodud interpreteerimise meetodeid paremaks, eriti filtrite läve leidmist. Samuti võiks DementiaBanki andmestikul püüda rakendada muid interpreteerimismeetodeid. Kindlasti võiks ka rakendada interpretatsioonimeetodeid muudel kliinilistel andmestikel või täiesti teistsugustesse domeenidesse kuuluvatel andmestikel.

Viidatud kirjandus

Aggarwal, C. C. & Zhai, C., 2012. *A Survey of Text Classification Algorithms.* Boston, MA, Springer, pp. 163-222.

Becker, J. T. et al., 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *Archives of Neurology*, 6, Volume 51, pp. 585-594.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, Volume 17, pp. 790-799.

Collobert, R. et al., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research 12*, pp. 2493-2537.

Fukunaga, K. & Hostetler, L. D., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Information Theory*, Volume 21, pp. 32-40.

Goldberg, Y., 2017. Neural Network Methods for Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 10(1), pp. 1-309.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning.* MIT Press.

Goodglass, H. & Kaplan, E., 1983. *The Assessment of Aphasia and Related Disorders.* Philadelphia: Lea & Febiger.

Hamming, R. W., 1950. Error detecting and error correcting codes. *The Bell system technical journal*, April, 29(2), pp. 147-160.

Harbecke, D., Schwarzenberg, R. & Alt, C., 2018. Learning Explanations from Language Data. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 316-318.

Hashimoto, K., Xiong, C., Tsuruoka, Y. & Socher, R., 2017. *A Joint Many-Task Model: Growing a Neural Network for Multiple NLP Tasks.* Copenhagen, Association for Computational Linguistics, pp. 446-456.

Jacovi, A., Shalom, O. S. & Goldberg, Y., 2018. Understanding Convolutional Neural Networks for Text Classification. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 21 November.pp. 56-65.

Kalchbrenner, N., Grefenstette, E. & Blunsom, P., 2014. A Convolutional Neural Network for Modelling Sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Volume 1, pp. 655-665.

Karlekar, S., Niu, T. & Bansal, M., 2018. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* pp. 701-707.

Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* pp. 1746-1751.

Kindermans, P.-J.et al., 2018. Learning how to explain neural networks: PatternNet and PatternAttribution. *International Conference on Learning Representations (ICLR).*

LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *nature*, 521(7553), p. 436.

Letarte, G., Paradis, F., Giguère, P. & Laviolette, F., 2018. Importance of Self-Attention for Sentiment Analysis. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* pp. 267-275.

MacWhinney, B., 2000. *The CHILDES Project: Tools for Analyzing Talk.*. 3rd ed. Mahwah(NJ): Lawrence Erlbaum Associates.

Mitchell, T. M., 1997. *Machine Learning.* 1st ed. New York: McGraw-Hill.

Pennington, J., Socher, R. & Manning, C., 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),* pp. 1532-1543.

Sirts, K., Piguet, O. & Johnson, M., 2017. Idea density for predicting Alzheimer's disease from transcribed speech. pp. 322-332.

Srivastava, N. et al., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research,* 15(1), pp. 1929-1958.

Turovsky, B., 2016. *Found in translation: More accurate, fluent sentences in Google Translate.* [Online] Available at: https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ [Accessed 9 May 2019].

Yancheva, M. & Rudzicz, F., 2016. Vector-space topic models for detecting Alzheimer's disease. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Volume 1, pp. 2337-2346.

Yang, Z. et al., 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480-1489.

Zeiler, M. D. & Fergus, R., 2014. Visualizing and understanding convolutional networks. *13th European Conference on Computer Vision, ECCV 2014*, pp. 818-833.

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Andreas Pung**,

(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **Konvolutsioonilisel neurovõrgul põhineva teksti klassifitseerimismudeli interpreteerimine kliinilisel andmestikul**,
   (*lõputöö pealkiri*)

   mille juhendaja on Kairit Sirts,

   (*juhendaja nimi*)

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi

Tartus, **10.05.2019**